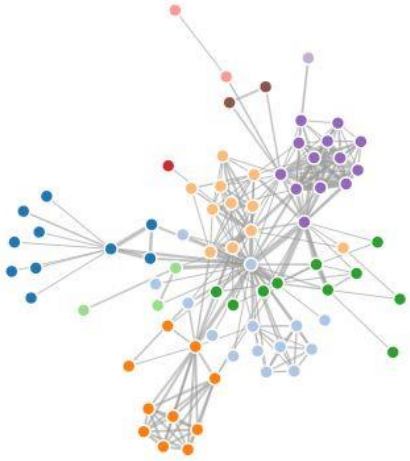
KSCHOOL

VI MÁSTER DATA SCIENCE

TFM



TÍTULO:

CONTAMINANTES INDUSTRIALES Y SALUD PÚBLICA

Autores:

David Zarco Yuste

Julián Antonio Guerrero Gijón

Curso 2017/2018

Indice

1 Introducción

Diversos estudios de distintas fuentes han relacionado ciertos contaminantes y su toxicidad con diversos tipos de enfermedades entre otras el Cáncer de pulmón, próstata o riñón.

La contaminación del aire es una amenaza para la salud mundial y causa millones de muertes humanas al año. Al no ser fácil de medir y cuantificar se hace necesario una mayor investigación y estudio de sus efectos y una política de protección mayor de los organismos frente a la salud humana.

1.1 Objetivo

El presente **Trabajo de Fin de Master** tiene como objetivo predecir mediante **técnicas de Data Science** una relación directa entre las emisiones Contaminantes de complejos Industriales ubicados en España y enfermedades por cáncer de pulmón codificada según el "**Manual de codificación CI-10-ES Diagnósticos"** elaborado por el Ministerio de Sanidad.

En el presente trabajo se definirán y modelarán diferentes variables, tanto del entorno como poblacionales, que puedan dar valor a nuestro proyecto y sirvan para establecer un modelo predictivo.

1.2 Generalidades

En proyecto se enfoca desde un punto de vista **generalista** y no se adentra es cuestiones médicas o de calidad de vida de la población, los cuales serían fundamentales para ese tipo de estudios.

Para el estudio se ha dado un papel importante a la **proximidad y distancia** al punto Contaminante, así como **tiempo de exposición** a este y la **Cantidad de contaminante** al que ha sido expuesto el sujeto fallecido.

El **tipo de Contaminante** es también otra variable de vital importancia al ser determinados compuestos de riesgo mayor que otros para la salud humana.

Hay que comentar que en el **modelo final** se utilizan únicamente dos tipos de Contaminantes ya que, en caso de haberse utilizado la totalidad, los resultados finales podrían enmascararse y no mostrar la realidad. Por ello y aunque en los ficheros de tratamiento de los datos se incluyen todos, se ha optado por elegir los contaminantes **Dióxido de Carbono CO_2 y Partículas PM10**.

Igualmente reseñar que después de procesar varios años de fallecimientos se ha cogido para los modelos sólo el **año 2015** por cuestiones de optimización y tiempo de procesado.

Como objetivo y con el fin de establece predicciones lógicas se ha cogido como "target" el fallecimiento producido por Tumor maligno de Pulmón (CI10- C34).

Entre otras variables también se ha incluido la **Edad** a la cual el sujeto fallece, aunque esta variable podría considerarse un **Factor de Confusión** ya que puede distorsionar la medida de la asociación entre otras dos variables.

En general para este tipo de proyectos sería necesario realizar un **estudio de cohorte** para comparar la frecuencia de la enfermedad entre dos poblaciones, así como estudios más detallados que incluyan la calidad de vida de la población y a nivel familiar.

Todos los datos en los cuales nos apoyamos son públicos y se han extraído de dos organismos independientes, estos se detallan más adelante.

El proyecto tiene un punto especial a considerar. Todos los datos del I.N.E. son abiertos pero los datos sobre fallecimientos de la población española con causa informada (motivo del fallecimiento) al ser muy sensibles, están sujetos a **condiciones especiales de protección** y sólo pueden utilizarse con fines de investigación para el presente estudio, por lo que estos no se subirán a github y sólo se expondrá una muestra.

1.3 Limitaciones del proyecto

Las limitaciones del proyecto son las siguientes:

- Se apoya en datos abiertos y no datos específicos recogidos para este tipo de estudios.
- Al no haberse desarrollado en Spark, nos hemos encontrado con limitaciones de recursos de máquina, habiendo procesos que requieren de tiempo de procesado (se indica más adelante para que no se vuelvan a recrear, comentándose el resultado final).
- Los fallecimientos con causa de muerte informada al no están georreferenciados y no estar registrados por domicilio conforme al sujeto, se han ubicado conforme a las coordenadas centrales de la población de residencia, esto puede hacer que las distancias no sean del todo reales, aunque si bien se ajustan en gran medida.

2 Estructura del Proyecto

En general, el proyecto tiene la siguiente estructura:

1. **Memoria explicativa del proyecto:** Este documento en formato PDF en el cual se describe el proyecto, el código desarrollado y las conclusiones.
2. **Repositorio GitHub**: Contiene el código y los datos necesarios para correr los scripts y notebook.

La URL del repositorio a consultar de **github** es:

https://github.com/dajuMasterDS/TFM_Industrials_Pollutants

2.1 Requisitos de instalación y librerías necesarias

Cualquier PC o pórtatil con al menos 8 GB de RAM será suficiente como para recrear el código. Todos los scripts que requieran para su ejecución una cantidad de tiempo superior al normal serán indicado en el código y en esta memoria.

Para correr determinados notebooks es necesario instalar algunas librerias Python específicas:

Para el análisis y modelos espaciales:

- **Pysal**: librería para análisis geoespacial *(#pip install pysal)*
- **Geopandas**: librería para tratamiento de datos georreferenciados *(#pip install geopandas)*
  #Dependencies: conda install -c conda-forge fiona shapely pyproj rtree

El proyecto se ha desarrollado tanto en R como en Python, para correrlo es necesario tener instalado **RStudio** y Anaconda con **Python 3**.

Existe un archivo de configuración del proyecto en R llamado Contaminantes.Rproj fácilmente ubicable a través del cual nos carga la configuración del proyecto en RStudio.

2.2 Estructura de carpetas

Dentro del repositorio se pueden encontrar las siguientes carpetas:

- Code: Contiene los scripts y notebook del estudio.

- Data: Contiene todos los datos de diferente tipo, tanto inputs, temporales y outputs.

La estructura es la siguiente:

	Carpetas		Descripción
code	scripts R y notebooks Python		
data	csv	Emissions	ficheros csv geocoding
		final	ficheros csv finales
		intermediary_csv	ficheros csv intermedios
		model_csv	ficheros csv para los modelos
		temp	ficheros csv temporales
	Deaths		Ficheros input INE
	doc		Manual CI-10 PDF
	Emissions		Ficheros xml Emisiones PRTR descomprimidos
	excel		ficheros excel auxiliares
	images		imágenes
	Industry		Ficheros xml Complejos PRTR descomprimidos
	shapes		shape files
	zip		Ficheros input zip files PRTR

3 Datos

Los datos en los cuales nos apoyamos son datos públicos y se han extraído de dos organismos independientes:

- Instituto Nacional de Estadística (INE): Se han obtenido los datos de fallecimientos de la población española entre los años 2012-2015. http://www.ine.es/

- Organismo estatal de Emisiones y Fuentes Contaminantes (P.R.T.R), de dónde se obtienen las fuentes contaminantes. http://www.en.prtr-es.es/

A continuación, se describe el proceso de obtención y clasificación de los datos.

3.1 Obtención de los datos

Los **ficheros con causa de muerte informada** proporcionados por el **INE** s están sujetos a condiciones especiales de protección y sólo pueden utilizarse con fines de investigación para el presente estudio y por las personas que realizan dicho estudio.

Estos se han obtenido en **formato zip** directamente por solicitud a este organismo, y pueden descomprimirse con el script de R 01-getINEData.R descargándose directamente a nuestra máquina en el directorio "../data/Deaths".

Estos datos contienen las defunciones registradas en España entre los años 2.012 al 2.015 con la causa de muerte codificada según el CI-10 y otros parámetros como edad defunción, municipio, ocupación, etc..

Se adjunta documento de condiciones de compromiso ES021-2018 Condiciones defunciones_completado.doc firmado con el INE en la carpeta "../data/doc/".

Los datos de **Contaminantes** del PRTR y los **Complejos industriales** causantes de estas, se obtienen directamente corriendo el script de R 02-getData_PRTR.R el cual descarga los ficheros zip al directorio "../data/zip". En total son 16 ficheros, 15 correspondientes a las emisiones producidas entre los años 2001 al 2015 y otro correspondiente a la descripción y ubicación de los Complejos industriales. Las ubicaciones no están georreferenciadas y sólo contienen la dirección y código postal asociado por lo que hay que Geocodificarlas en un primer paso. Cada complejo está asociado a un Código PRTR y a su vez puede ser fuente contaminante de varios Compuestos. En los ficheros de

Emisiones se describe el tipo de Contaminante asociado a cada Complejo (código PRTR), medio al que se vierte, cantidad emitida y otras variables.

3.2 Descripción de los datos

Los datos input y su descripción son como se muestran a continuación (en color los de interés):

3.2.1 Datos complejos industriales:

Campo	Descripción
CodigoPRTR	Código PRTR
NombreDelComplejo	Nombre complejo
EmpresaMatriz	Empresa matriz
ActividadEconomica	Actividad economica
CNAE-2009	Código CNAE-2009
CodPRTRDEI	Código PRTR DEI
CodIPPC	Código PPC
Direccion	Direccion
CodPostal	Código Postal
Poblacion	Población
Municipio	Municipio
Provincia	Provincia
CCAA	Comunidad Autónoma
DemarcacionHidrografica	Demarcación Hidrográfica

3.2.2 Datos emisiones industriales:

Campo	Descripción
CodigoPRTR	Código complejo
NombreDelComplejo	Nombre
AnyoReferencia	Año referencia
Contaminante	Contaminante
CantidadTotalkgporaAño	Cantidad total kg año
MedioReceptor	Medio receptor
Metodo	Método
Metodo_MCE	Método MCE

3.2.3 Datos fallecimiento INE:

Los ficheros de defunciones en España facilitados por el INE, de los años 2012 a 2015 (un fichero por cada año), contienen la causa de muerte desglosada en 4 variables: causa de muerte base 1, 2, 3 y causa de muerte perinatal.

Los valores de las variables causantes de muerte responden a la codificación CIE-10-ES del Ministerio de Sanidad.

Estos datos al ser datos en formato texto plano (microdatos) y estar codificados tienen una estructura diferente. A continuación, se muestra un ejemplo de ellos:

```
0105910201511020151108101059  101059  1 0000000666331P2990822834990
2807911201561120151108128079  128079  1 0000000666331Q9990858247990
2807912201511220151108128079  128079  1 0000000666331P9690824939990
3907512201511220151108139075  139052  1 0000000663331P2400822434990
4718612201511220151108147186  147186  1 0000000666331P0710821632990
```

El formato del fichero y valor de las variables, excepto causas de muerte, se recoge en el documento:

"../data/excel/diseño_registro_anonimizado_cm_nivel_estudios2015.xls"

En este fichero se describe la posición que ocupa cada variable y su codificación.

Un ejemplo de la descripción de los primeros registros es esta:

Variable	Inic.	Fin.	Long.	Descripción	Tipo de campo - Valores válidos
CPROI	1	2	2	Código Provincia de Inscripción	Numérico entre 01 y 52 e igual al código de la provincia en la que se está grabando.
CMUNI	3	5	3	Código Municipio de Inscripción	Numérico, compatible con diccionario geográfico. No se admiten "000", "999" y " ".
MESN	6	7	2	Mes-fecha de nacimiento	Numérico, entre 01 y 12.
ANON	8	11	4	Año-fecha de nacimiento	Numérico.
SEXO	12	12	1	Sexo	Numérico: 1=varón, 6= mujer.
MESDEF	13	14	2	Mes-fecha de la defunción	Numérico, entre 01 y 12.

El fichero de defunciones con causa de muerte informada está bajo contrato de confidencialidad.

3.3 Datos auxiliares

En el estudio nos hemos utilizados datos auxiliares, estos son:

- Manual internacional de codificación de enfermedades CI-10 publicado por el Ministerio de Sanidad.
  Ubicado en la ruta "../data/doc/UT_MANUAL_DIAG_2016_prov1.pdf".
- Libro Excel con las **geolocalizaciones de los municipios de España**. Contienen el nombre de la provincia y municipio. No contiene el código numérico de municipio Se ubica en la ruta "../data/Deaths/MunicipiosGeolocalizados.csv".
- Shape files extraidos del Instituto Geográfico Nacional, ubicados en la ruta "../data/shapes/".
- Fichero Excel con la relación de nombres de provincia y municipios españoles con códigos numéricos de identificación de provincias y municipios. Se ubica en la ruta "../data/Deaths/10codmun.xls".
- CIE-10.csv: listados causas de muerte CIE-10 por primera causa de fallecimiento.
- Libro Excel con las codificaciones de 3 y 4 dígitos y su descripción. Ubicado en la ruta "../data/excel/CIE10_10rev.xlsx".

4 Limpieza y preparación del dato

En este capítulo se describe el proceso de limpieza y preparación del dato.

4.1 Obtención y preparación del dato: Fase 1

Esta fase es la primera del proyecto y se ha desarrollado en R. Comprende los Scripts:

Script R
01-getINEData.R
02-getData_PRTR.R
03-GeolocateIndustryByEmissions.R
04-GeolocateINEData.R

El **objetivo** de la fase es generar un archivo único de Complejos industriales Georreferenciados asociados a sus Emisiones durante los años 2.001-2.015.

El dato al ser obtenido de dos fuentes diferentes se trata por separado. Una vez analizadas e identificadas las variables que son de interés para el estudio se van recogiendo y pre-procesando. Debido a que los complejos no están ubicados por Longitud/Latitud se hace necesaria su geocodificación.

A continuación, una descripción de los Scripts:

- Script **01-getINEData.R**: Este fichero descomprime y lee cada uno de los ficheros fuente facilitados por el INE en txt, con las defunciones en territorio español y causas de muerte de 2012 a 2015, dando formato a las distintas variables e incluyendo separador ';' para generar finalmente un único fichero "DfDefuncionesSinGeo.csv".
- Script **02-getData_PRTR.R**: Este fichero obtiene de las URL de origen los ficheros zip comprimidos y los descomprime en la estructura de directorios. Posteriormente los descomprime y almacena en su formato original (XML).
- Script **04-GeolocateINEData.R**: Este fichero prepara el fichero de defunciones, incluyendo la longitud-latitud de cada uno de los fallecimientos, a nivel de municipio, imprescindible para relacionar más adelante con los registros de emisiones contaminantes.

En este script se genera el dataframe de códigos numéricos de municipios y geolocalización de municipios a partir de los ficheros fuentes auxiliares listado-longitud-latitud-municipios-espana.xls y 10codmun.xls. El resultado se guarda en fichero "../data/Deaths/DfDefuncionesGeo.csv" para ser utilizado en próximos pasos.

- Script **03-GeolocateIndustryByEmissions.R** Este fichero, obtiene los archivos de Complejos Industriales y Emisiones por Complejo de la estructura de directorios y lee el formato XML, generando dos Dataframes, uno con el total de emisiones por años y otro con los Complejos industriales. Posteriormente hace un merge por el código PRTR y preparar ligeramente los datos para su geocodificación posterior.

Como generalidades a destacar, hay que comentar que el proceso de obtención y descarga no tiene dificultad alguna pero no así el proceso de **Geocodificación** el cual se ha realizado llamando al API de Google mediante la función geocode del bloque 5 del script *03-GeolocateIndustryByEmissions.R (Bloque de geocodificación de los complejos*).

Hay que destacar que el máximo de llamadas diarias al API es de 2.500 peticiones y por otro lado **mala calidad** de las direcciones de ubicación de los Complejos del PRTR.

Muchas llamadas no obtienen resultado debido al estado del dato, por lo que ha sido necesario limpiar este e intentar la geocodificación por direcciones diferentes.

El resultado de la llamada es un par de coordenadas geográficas WGS84 "Longitud" / "Latitud" en formato decimal.

Debido a que el número de llamadas al API es limitado y en una pasada "consume" las request, es necesario ir generando CSV temporales para luego reconstituirlos en el Bloque 6 *(Bloque de generación de outputs y unión de ficheros*).

El paso siguiente es el bloque 7 (*Reconstitución del dataframe y fichero final*) dónde se reconstituyen los CSV parciales a uno final.

Posteriormente y en un paso último, mediante filtrado manual en Excel se han limpiado las direcciones con valores de Latitud/Longitud erróneas fueran del rango del territorio español. Estas direcciones se han desechado.

El resultado final es **5.006 complejos Industriales geolocalizados** de una total inicial de 7.000.

Los **ficheros resultantes son dos:**

- Fichero CSV (convertido a Excel por conveniencia), conteniendo los Complejos industriales georreferenciados asociados a su emisión por tipo de contaminante durante los años 2.001 al 2.015. Se ubican en la ruta "../data/csv/Emissions/geo/Comp_by_Emi_total_Geo_2001-2015.xlsx".
- Fichero CSV con las defunciones almacenadas según sus valores y georreferenciadas. Se ubican en la ruta "../data/Deaths/DfDefuncionesGeo.csv"

Variables recogidas ficheros INE					
ProvinciaReside	AnioDefuncion	PaisResidencia	TamanioMuniResi	CausaMortaReduc	Longitud
MunicipioReside	Nacionalidad	EstadoCivil	TamanioPaisNaci	CausaMortaperin	Altitud
ProvinciaInscri	PaisNacimiento	Ocupacion	TamanioPaisResi	CausaMortaInfan	Habitantes
MunicipioInscri	LugarNacimiento	AnioCumplidos	TamanioPaisNdad	NivelEstudios	Hombres
MesNacimiento	ProvNacimiento	MesesCumplidos	CausaMuertebas1	PoblaciÃ³n	Mujeres
AnioNacimiento	MunicipioNacimi	DiasCunmplidos	CausaMuertebas2	Provincia	codpostal
Sexo	PaisNacimiento	TamanioMuniInsc	CausaMuertebas3	Comunidad	
MesDefuncion	LugarResidencia	TamanioMuniNaci	CausaMuertebas4	Latitud	

Variables recogidas ficheros PRTR	
CodigoPRTR	CodPRTRDEI
NombreDelComplejo1	CodIPPC
NombreDelComplejo2	Direccion
EmpresaMatriz	CodPostal
ActividadEconomica	Poblacion
AnyoReferencia	Municipio
Contaminante	Provincia
CantidadTotalkgporaAnyo	CCAA
MedioReceptor	DemarcacionHidrografica
Metodo	Latitud
Metodo_MCE	Longitud
CNAE-2009	

4.2 Preparación del dato: Fase 2

Esta fase del proyecto está enteramente desarrollada en **Python**, a través de notebooks.

4.2.1 Unión de ficheros de Defunciones con fichero Emisiones Contaminantes

El notebook "**05-MergeDeathsEmissions.ipynb"** realiza la unión del fichero geolocalizado de defunciones"DfDefuncionesGeo.csv" con el fichero geolocalizado de emisiones contaminantes "Comp_by_Emi_total_Geo_2001-2015.xlsx" . Esto se realiza antes de la unión de ficheros defunciones-emisiones contaminantes, se eliminan del estudio los fallecimientos por algunas causas de muerte que por definición no parecen tener una causa directa con las emisiones contaminantes, estas son:

a. Enfermedades infecciosas y parasitarias A00-B99
b. Malformaciones congénitas, deformaciones y anomalías cromosomáticas Q00-Q99.
c. Lesiones y envenenamientos S00-S99
d. Factores Sanitarios Z00-Z99

Se eliminan, igualmente, todas aquellas emisiones que no se han propagado por el aire (variable "Medio Receptor").

Se genera un área de influencia alrededor de la posición del fallecimiento con nuevas variables Lontidud1 y 2 Latitud1 y 2 que describen el área cuadrada sobre la que actuará la emisión del contaminante para cada fallecimiento.

Para cada defunción se genera un registro por emisión de contaminante a la que ha estado expuesto y año de emisión, desde su nacimiento hasta el año de defunción, dentro del área cuadra calculada.

Graba fichero de salida a la ruta "../data/csv/merge" con nombre "DeathsEmissionsFinal_AAAA_NNN.csv", cada 500 emisiones. Estas son tratadas así para liberar memoria, donde AAAA es el año de la emisión y NNN el número de emisiones tratadas para ese año.

El fichero de salida contiene el formato del fichero de defunciones más las variables código PRTR, año de referencia, contaminante, cantidad total de kg, coordenadas del Dataframe de contaminantes.

4.2.2 Agrupación de ficheros Fallecimientos-Emisiones por Contaminante.

El Script **"06-GroupDeathsEmissions.ipynb"** trata todos los ficheros generados en el script anterior y los agrupa por contaminante para su tratamiento posterior, a partir de una lista de contaminantes únicos cuyo medio de emisión es el aíre.

Graba fichero de salida cada 1.500.000 para liberar memoria.

El nombre del fichero generado es "DeathsEmissions_NCONT_CONT.csv", donde NCONT hace referencia al literal del contaminante y CONT a la secuencia de fichero generado para un mismo contaminante.

La ruta de almacenamiento es "../data/csv/group"

4.2.3 Ficheros de ficheros Fallecimientos-Contaminantes totales por año 2015 y totales acumulados por Cantidad Contaminante y tiempo exposición.

El Script **"07-FinalDeathsEmissions.ipynb"** vuelve a procesar los ficheros generados por contaminante para agrupar cada una de las defunción-contaminante-año poraño de emisión (en este caso 2015) en un sólo registro, creando nuevas variables total años de exposición y total kg contaminante.

Graba fichero con nombre "DeathsEmissions_final_NCONT_1.csv", donde NCONT hace referencia al nombre del contaminante. La ruta de almacenamiento es "../data/csv/final"

Nota*: Para los modelos propuestos en los apartados siguientes se tomarán únicamente dos Contaminantes para recrear los modelos. Estos son:

#DeathsEmissions_final_Dioxido_de_carbono_CO2_1.csv

#DeathsEmissions_final_Particulas_PM10_1.csv

Debido a que los notebooks que siguen recogen los CSV por Contaminante para montar el Dataframe final, los Contaminantes de interés se indicarán por el símbolo "#" al inicio del fichero.

5 Generación del dataframe para el modelo.

El notebook **08-GenDataframeModel.ipynb** genera el dataframe para el modelo final para que posteriormente se puedan aplicar las librerías espaciales para la generación de distancias.
El archivo realiza un merge con el archivo de enfermedades codificado "../data/excel/CIE10_10rev.xlsx' y establece el campo "**target**" para el tipo de enfermedad C34-CI10 por tumor maligno de Pulmón.

Finalmente genera el fichero '../data/csv/model_csv/df_model.csv' el cual será utilizado posteriormente.

6 Análisis de datos y Visualización

Para el análisis de datos se han desarrollado dos notebooks de Python.

6.1 Análisis de Contaminantes

En el notebook **"09-Analysis_Emissions.ipynb"** se analizan distintas emisiones asociadas a Contaminantes. Los datos utilizados para esta visualización son los obtenidos del fichero cruce de Emisiones-Complejos y no contienen información sobre fallecimientos.

Los gráficos generados nos dan una perspectiva global de los tipos de Contaminantes y su distribución.

En general la tabla de frecuencias relativas nos da información acerca de los porcentajes de Emisiones y el medio en el cual son vertidos, vemos que las **emisiones al "Aire**" son las más frecuentes, esas serán objeto de nuestro estudio.

Aire	74.652651
Litoral	9.504215
EDAR de titularidad pública (municipal o autonómica)	3.276913
Cuenca intercomunitaria de titularidad estatal	3.235265
Cuenca intracomunitaria de titularidad autonómica	2.535568
Red de alcantarillado	2.420618
EPER Cuenca Intercomunitaria	1.187819
Cuenca intracomunitaria	0.653050
Depuradora privada externa al complejo industrial	0.609736
Red de alcantarillado sin depuración (municipales o autonómicos)	0.599740
cuenca intercomunitaria gestionada por comunidad autónoma	0.581415
EPER Depuradora	0.516443
Cuenca intracomunitaria de titularidad estatal	0.226568

Tabla 1

Si contabilizamos los porcentajes relativos de Contaminantes emitidos por Kgs al Aire durante los años 2001-2015 vemos que el **Amoniaco (NH3)** es el más numeroso en cuanto a emisiones, este fundamentalmente proviene de industria asociada a granjas. Para nuestro estudio vamos a obviar este entre otros por considerarlo uno de los menos nocivos para la salud.

Porcentajes relativos de cantidades de contaminantes emitidos 2001-2015 (Kgs)	
Amoniaco (NH3)	44.638593
Óxidos de nitrógeno (NOx/NO2)	9.600321
Dióxido de carbono (CO2)	5.862400
Metano (CH4)	5.007699
Óxidos de azufre (SOx/SO2)	4.213251
Partículas (PM10)	3.601794
Compuestos orgánicos volátiles distintos del metano (COVNM)	3.010422
Níquel y compuestos (como Ni)	2.793957
Flúor y compuestos inorgánicos (como HF)	2.323090
Monóxido de carbono (CO)	2.191426
Zinc y compuestos (como Zn)	2.012899
Cloro y compuestos inorgánicos (como HCl)	1.653612
Óxido nitroso (N2O)	1.646917
Plomo y compuestos (como Pb)	1.537569
Mercurio y compuestos (como Hg)	1.519716
Cadmio y compuestos (como Cd)	1.459463
Arsénico y compuestos (como As)	1.138113
Cromo y compuestos (como Cr)	1.001986
Benceno	0.865859
Cobre y compuestos (como Cu)	0.803374
Hidrocarburos aromáticos policíclicos totales PRTR (HAP totales PRTR)	0.787753
PCDD + PCDF (dioxinas + furanos) (como Teq)	0.493182
Hidrofluorocarburos (HFC)	0.325813
Cianuro de hidrógeno (HCN)	0.252170
Naftaleno	0.185222
Hidroclorofluorocarburos (HCFC)	0.176296
Ftalato de bis (2-etilhexilo) (DEHP)	0.162906
Diclorometano (DCM)	0.160675
Perfluorocarburos (PFC)	0.087032
Tetracloroetileno (PER)	0.066948
Cloruro de vinilo	0.062485
Policlorobifenilos (PCB)	0.060253
Tricloroetileno	0.046863
Tetraclorometano (TCM)	0.042400
Triclorometano	0.040169
Clorofluorocarburos (CFC)	0.033474
1,2-dicloroetano (DCE)	0.026779
Hexaclorobenceno (HCB)	0.024548
Antraceno	0.022316
Pentaclorofenol (PCP)	0.015621
Triclorobencenos totales (TCB)	0.011158
Óxido de etileno	0.008926
1,1,1-tricloroetano (TCE)	0.008926
1,1,2,2-tetracloroetano	0.006695
Hexafluoruro de azufre (SF6)	0.006695
1,2,3,4,5,6-hexaclorociclohexano (HCH)	0.002232

Tabla 2

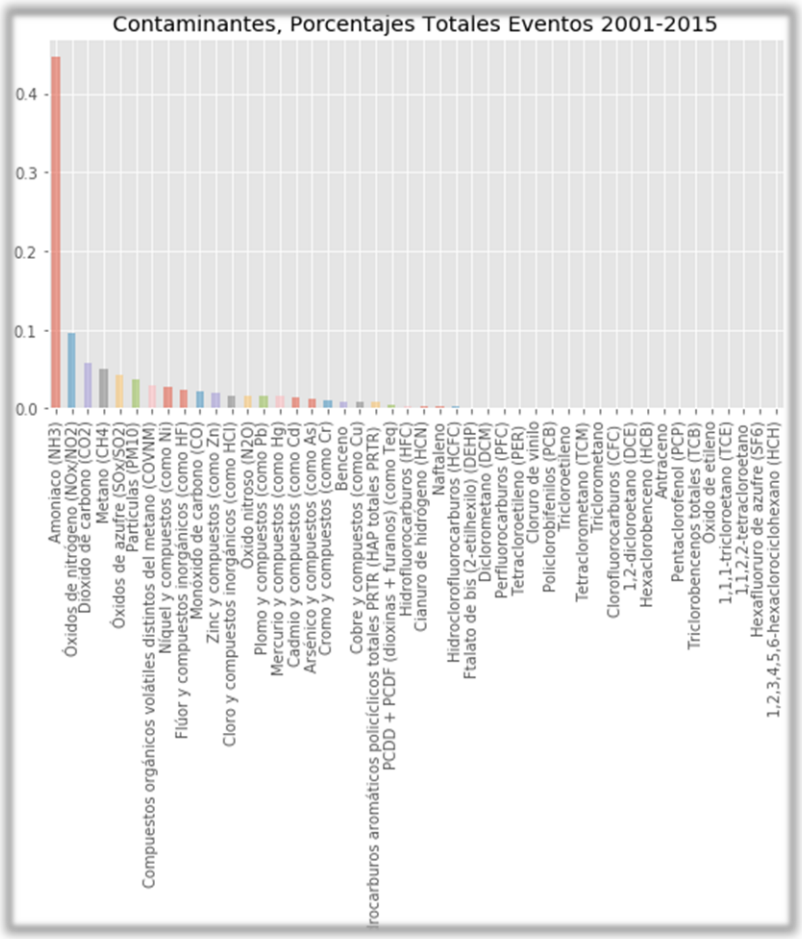


Figura 2

La figura 3 muestra los 10 Contaminantes no por frecuencia sino por Kgs de emisión emitidos a la atmosfera, se observa que el CO_2 es el más importante seguido por los Óxidos de azufre.

Contaminante	Total_Kg
Dióxido de carbono (CO2)	2.006112e+12
Óxidos de azufre (SOx/SO2)	9.020753e+09
Óxidos de nitrógeno (NOx/NO2)	5.632728e+09
Monóxido de carbono (CO)	3.798864e+09
Metano (CH4)	2.445293e+09
Compuestos orgánicos volátiles distintos del metano (COVNM)	8.796702e+08
Amoniaco (NH3)	5.828544e+08
Partículas (PM10)	3.823163e+08
Óxido nitroso (N2O)	8.338749e+07
Cloro y compuestos inorgánicos (como HCl)	3.509602e+07

Figura 3

Visualización del total de Contaminantes en el periodo 2001 a 2015, existe una tendencia en ascenso hasta el año 2007 a partir del cual comienza a descender ligeramente para luego repuntar en el 2014.

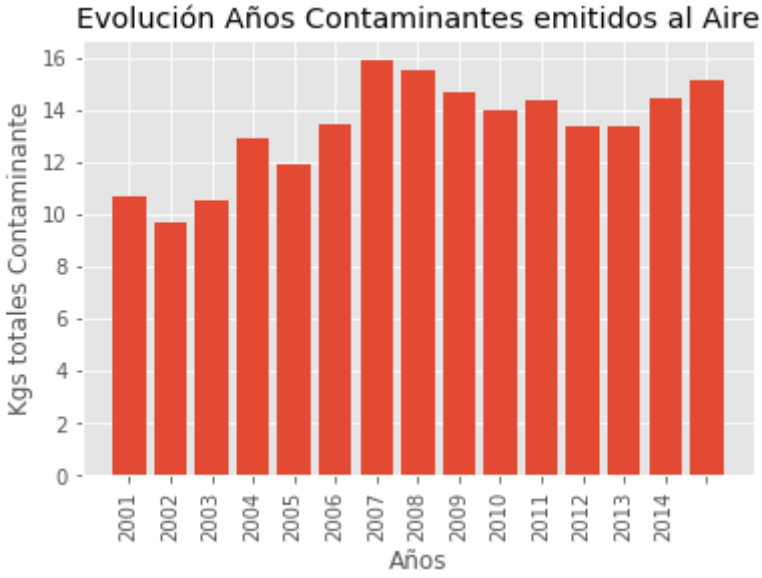


Figura 4

Para nuestro estudio tomaremos en cuenta las emisiones de CO_2 así como las partículas M10 y trataremos de ver las relaciones con causas de fallecimientos provocados por Cáncer de pulmón (CI10- C34).

6.2 Análisis de defunciones vinculadas a fuentes contaminantes

El notebook "**10-AnalysisDeathsvsEmissions.ipynb"** genera un gráfico total defunciones, para cada tipo de muerte, por contaminante.

Estos gráficos se exponen para el Dióxido de carbono (CO_2) y Partículas (PM10), los cuales han sido elegidos para el estudio.

Contaminante CO_2

Partículas PM10

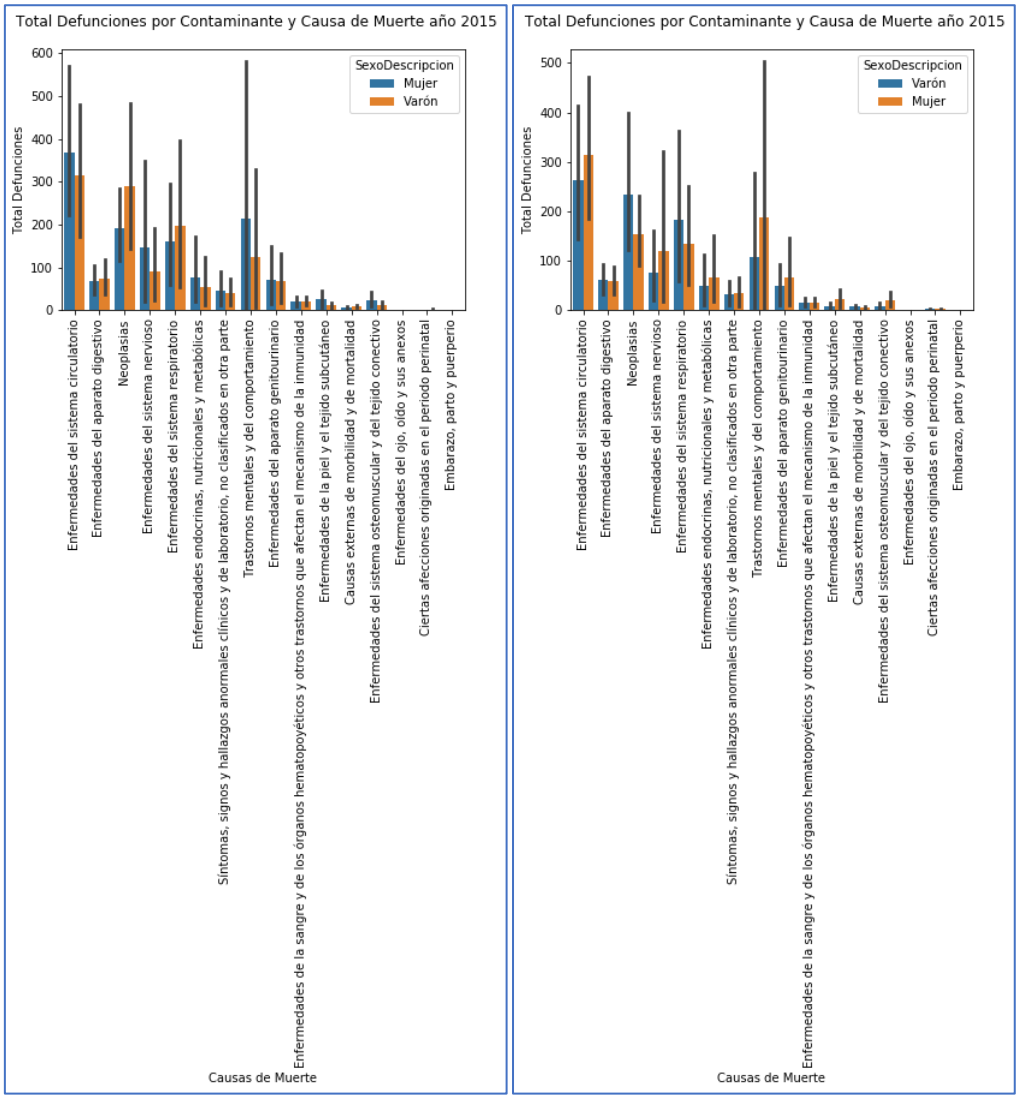
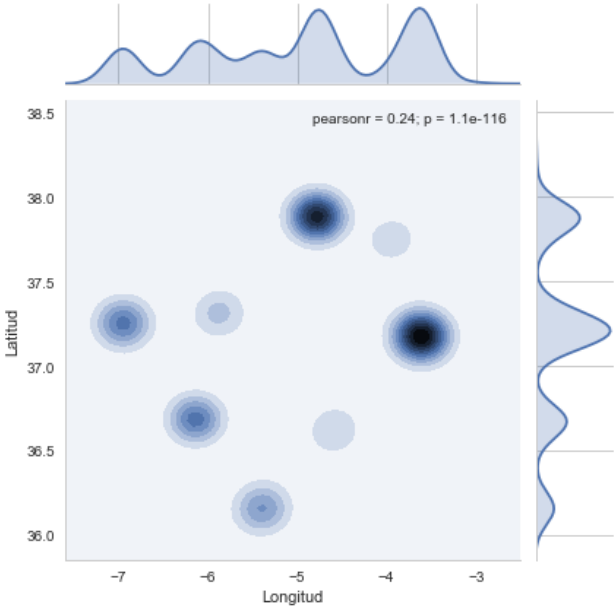


Figura 1

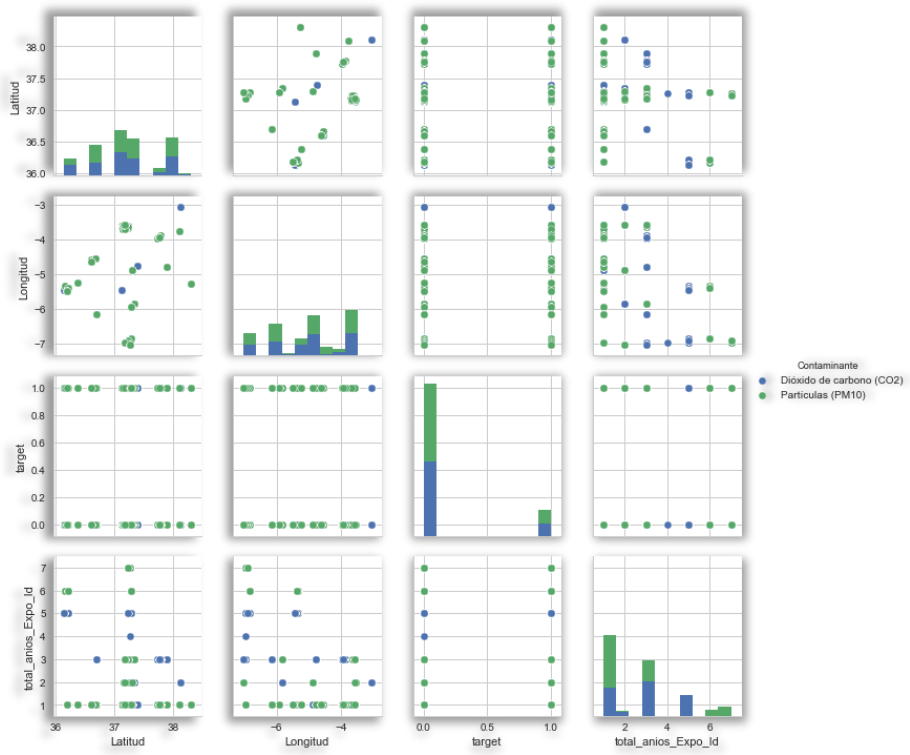
En los gráficos pueden verse distintas asociaciones de defunciones por tipo según el Contaminante elegido.

En el notebook **11-AnalysisPre-modelVisualization.ipynb** se realizan diversas visualizaciones recogidas del fichero '../data/csv/model_csv/df_model_geom.csv' generado a través del notebook **12-Geometry.ipynb** el cual se explicará más adelante (este último genera distancias al punto contaminante).

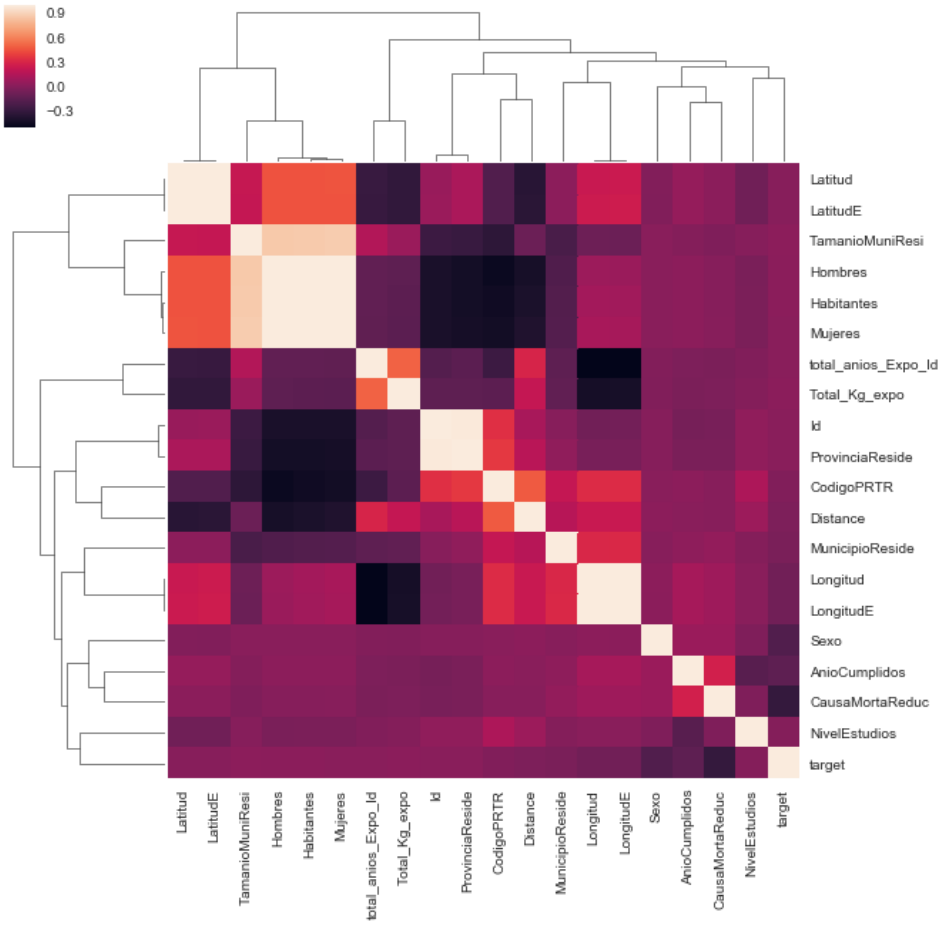
En el fichero se hacen visualizaciones varias y se concreta una zona de estudio por razones de optimización, para plotear datos de la Comunidad Autónoma de Andalucía. La gráfica siguiente muestra un jointPlot con las variables Latitud y Longitud, en la que puede observarse las concentraciones de fallecimientos acotados en torno al foco de contaminación.



En el gráfico siguiente se realiza un pairplot por el Contaminante (CO_2 y PM10) ploteando las variables Latitud y Longitud fallecimientos, total años exposición y target, puede extraerse relaciones según el tipo de Contaminante y la variable a considerar.



Abajo podemos ver un **clustermap** con distintas variables consideradas:



7 Análisis geoespacial y generación geometrías

El notebook **12-Geometry.ipynb** genera geometrías tipo “POINT” de los eventos de fallecimientos y focos contaminantes y calcula las distancias entre estos dos puntos.

Para ello es necesario tener instaladas ciertas librerías de datos geoespaciales comentadas anteriormente en el apartado 2.1 como son Geopandas y Shapely.

El fichero convierte previamente las coordenadas geográficas WGS84 a coordenadas proyectadas UTM para la zona 30N, requisito necesario para calcular las distancias.

El script salva a CSV el fichero '../data/csv/model_csv/df_model_geom.csv' con las geometrías generadas.

Finalmente se salva a shape file, el cual se utiliza para realizar visualizaciones.

8 Modelos propuestos

En este apartado se describen las metodologías y técnicas estadísticas propuestas para relacionar las emisiones con los fallecimientos.

En nuestro estudio se ha optado por aplicar las siguientes técnicas de machine learning:

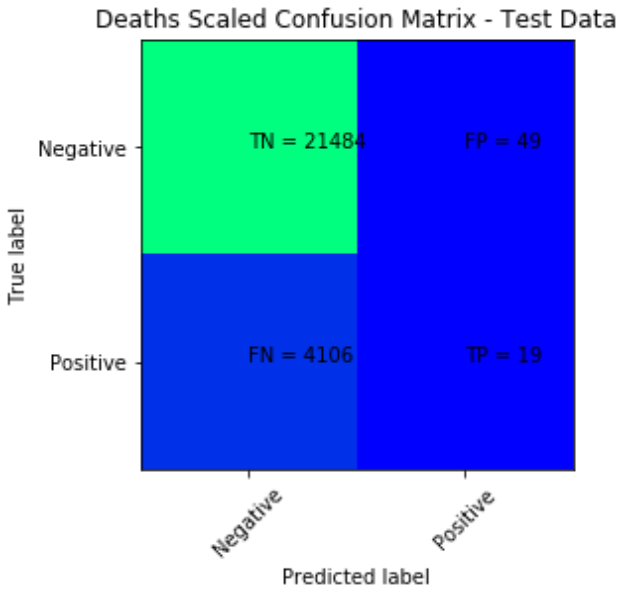
- Knn Neighbors.
- Spatial Regression, importante para ver la no-independencia espacial.
- Comparación de modelos GaussianNB, DecisionTreeClassifier, LogisticRegression y KneighborsClassifier.

8.1 Knn Neighbors

Knn Kenighbors es un método de clasificación no paramétrico, el cual hemos elegido ya que estima la probabilidad a posteriori de que un elemento pertenezca a una clase a partir de la información proporcionada por el conjunto. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

En el notebook **13-Knn-Neighbors.ipynb** se describe todo el modelo y los resultados.

Hay que comentar que este modelo arroja un “Accuracy” medio de 0.827



8.2 Spatial Regression

En el notebook **14-Model_Spatial_Regression.ipynb** se describe el modelo de Autocorrelación espacial realizado junto con un modelo de regresión lineal por MCO.

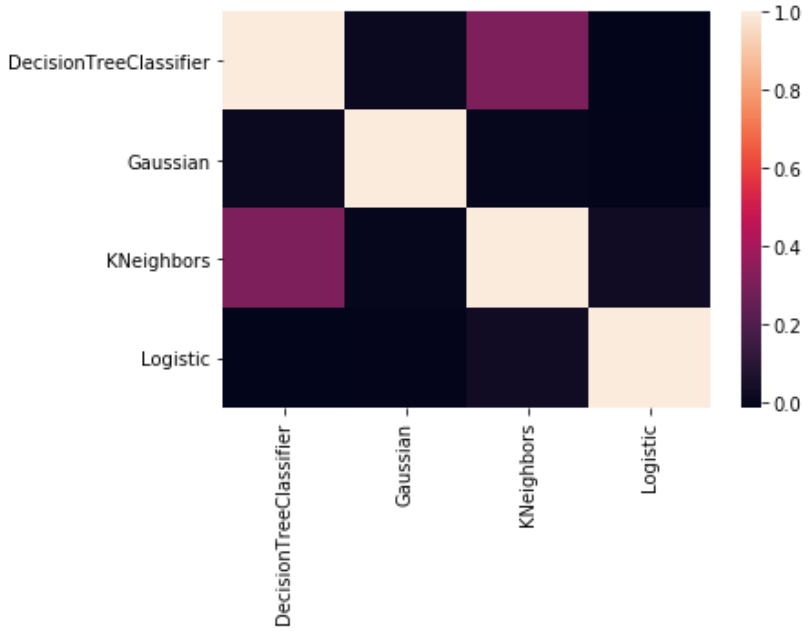
8.3 Ensemble Models: GaussianNB, DecisionTreeClassifier, LogisticRegression y KneighborsClassifier

En este último script "**15-ModelsPredDeathsEmissions**" se comparan los modelos GaussianNB, DecisionTreeClassifier, LogisticRegression y KneighborsClassifier y aplica varios metaclasificadores:

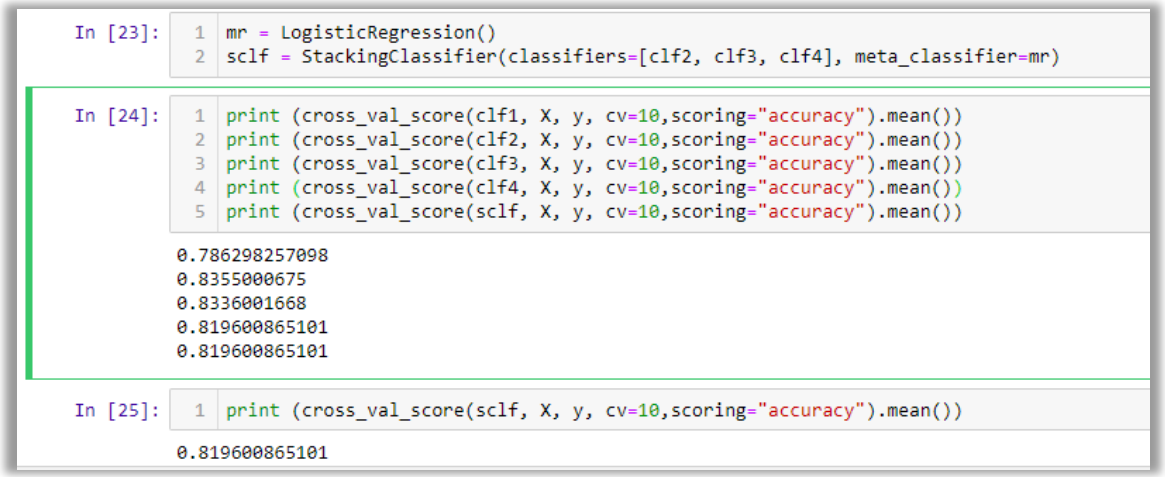
- VotingClassifier.
- StackingClassifier
- BaggingClassifier
- XGBClassifier

En este modelo después de establecer el set de datos de entrenamiento y test, se utilizan 4 clasificadores diferentes para obtener el "accuracy" medio.

Después de entrenar el modelo con estos clasificadores se hace una predicción cuya correlación que arroja lo siguiente:



Posteriormente utilizamos varios metaclasificadores variando los modelos utilizados en origen, un ejemplo es el Stacking dónde utilizamos 3 modelos como son el Gaussian, Logistic y KNeighbors. El resultado es un accuracy medio como muestra la imagen:



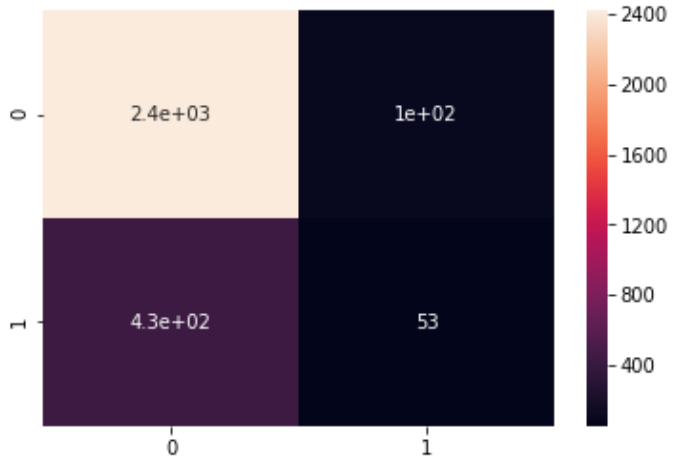
```
In [23]:
mr = LogisticRegression()
sclf = StackingClassifier(classifiers=[clf2, clf3, clf4], meta_classifier=mr)

In [24]:
print (cross_val_score(clf1, X, y, cv=10,scoring="accuracy").mean())
print (cross_val_score(clf2, X, y, cv=10,scoring="accuracy").mean())
print (cross_val_score(clf3, X, y, cv=10,scoring="accuracy").mean())
print (cross_val_score(clf4, X, y, cv=10,scoring="accuracy").mean())
print (cross_val_score(sclf, X, y, cv=10,scoring="accuracy").mean())

0.786298257098
0.8355000675
0.8336001668
0.819600865101
0.819600865101

In [25]:
print (cross_val_score(sclf, X, y, cv=10,scoring="accuracy").mean())

0.819600865101
```

En este modelo se ha utilizado también otros metaclasificadores cuyos resultados medios en el “Accuracy” son los siguientes:

Metaclasificadores	Mean Accuracy
VotingClassifier	0.825900266
StackingClassifier	0.819600865
BaggingClassifier	0.835500068
XGBClassifier	0.823666667

Generamos los estadísticos del metaclasificador XGBClassifier y los ploteamos en su matriz de confusión.

	precision	recall	f1-score	support
0	0.85	0.96	0.9	2520
1	0.34	0.11	0.17	480
avg / total	0.77	0.82	0.78	3000



En conclusión podemos decir que el mejor Modelo para implementar según el Accuracy lo obtiene el metaclasificador BaggingClassifier con el estimador GaussianNB.

9 Conclusiones

Las conclusiones a las cuales hemos llegado en el estudio son las siguientes:

- Es necesario focalizar más el estudio a zonas concretas, causas de fallecimientos y tipos de Contaminantes. Los modelos no pueden generalizar correctamente.
- La **edad** como variable en el modelo es un factor que hay que considerarla como factor de confusión ya que esta puede estar asociada a factores naturales relacionados con el envejecimiento.
- Tanto las variables **tiempo de exposición** al contaminante como **total de Kgs exposición** son variables que influyen considerablemente en los casos de fallecimiento por cáncer.
- La **distancia** al foco contaminante es una variable decisiva en el modelo ya que a menores distancias al foco la probabilidad de fallecimiento por cáncer de pulmón es mayor.
- La georreferenciación de los fallecimientos debería realizarse por domicilio y no por municipio para establecer una relación más directa con la emisión contaminante.
- Para estudios posteriores se deberían incluir variables del entorno naturales como la dispersión de contaminantes en el aire, así como factores poblacionales en relación con la calidad de vida de la población y la zona. Las variables sociosanitarias serían también muy importantes a tener en cuenta para este tipo de estudios (hábitos alimenticios, consumo tabaco-alcohol, etc.).

Referencias

[1] Instituto de Salud Carlos III de Madrid http://www.isciii.es/

[2] Instituto Salud Global de Barcelona. https://www.isglobal.org/

[3] OMS http://www.who.int