KSCHOOL

INCENDIOS EN GALICIA

JAVIER RUIBAL FERNÁNDEZ
MÁSTER EN DATA SCIENCE

Contenido

Motivación

Me gustaría empezar este trabajo con una frase de una de las personas más reconocidas de la literatura gallega. Cuando eres consciente y valoras lo que tienes, y aun así tienes que marcharte de Galicia, te invaden la *morriña* y las ganas de volver. Es duro, año tras año, ver en las noticias como arde Galicia, y más duro aún es ir por el coche por donde solías ir, y encontrarte con que a tu alrededor ya no hay bosque, sino tierra y cenizas.

Espero con este proyecto poder aportar mi pequeño granito de arena a la lucha contra incendios en la que tantos profesionales están involucrados.

Introducción

En el presente trabajo de fin de máster se pretende, además de dar visualización a la cantidad de incendios ocurridos en Galicia, realizar un estudio exhaustivo de los diversos incendios acontecidos entre 2001 y 2015. A lo largo de este documento, se expondrán diferentes técnicas de machine learning utilizadas para optimizar el control de futuros incendios.

Como primer paso, en el capítulo 1, detallaremos la limpieza de datos aplicada a la tabla principal, posteriormente realizaremos un análisis descriptivo de los datos y, por último, proporcionaremos información adicional relacionada con el clima y la geografía gallegas que servirán de utilidad para futuros estudios.

Los datos obtenidos en el Capítulo 1 nos proporcionan la información de los incendios ocurridos en Galicia en la ventana temporal mencionada. Sin embargo, para utilizar modelos de clasificación, en el Capítulo 2 hemos decidido realizar un análisis mensual y crear una nueva tabla donde, además de contener los puntos afectados de cada mes, se pueden observar los focos que no se han detectado incendiados durante el mismo.

El Capítulo 3 consiste en un estudio de la geolocalización de las actuales torres de control ubicadas en Galicia. En función del histórico disponible de incendios, realizaremos una nueva distribución para las 44 torres de control actuales, para ello utilizamos el método de K-Means adaptado a distancias geodésicas.

Por último, en el Capítulo 4 presentaremos un front-end donde se exponen tres pestañas: la principal donde el usuario podrá observar un análisis de los datos en función del año y provincias deseadas; la segunda pestaña se corresponde con la utilización del modelo de predicción definido, aquí el usuario seleccionará un municipio de Galicia para conocer la probabilidad de incendio en función de los últimos datos registrados por la estación meteorológica más cercana; y por último, en la tercera pestaña, el usuario podrá realizar una comparación entre la localización de las actuales torres de control con las torres obtenidas mediante nuestro modelo.

Capítulo 1

Extracción y limpieza de datos

Para realizar este trabajo partimos inicialmente de una base de datos de los incendios ocurridos en España a lo largo de los años 2001 a 2015 proporcionada por el portal de datos CINVIO (Ministerio de Agricultura, s.f.).

A continuación, proporcionamos la visualización de la tabla inicial a alto nivel:

id	superficie	fecha	idprovincia	...	time_ctrl	time_ext	personal	medios	perdidas
2001150021	5.00	20/02/2001	15	...	235	270	14	2	7013
2001150088	1.50	24/02/2001	15	...	470	530	14	1	1497
2001150090	3.00	25/02/2001	15	...	185	220	14	3	1882
2001150094	1.50	25/02/2001	15	...	125	135	5	0	1028
2001150111	3.80	25/02/2001	15	...	1050	1051	14	1	3119
...	...	...	...	...	...	...	...	...	...
2015360747	1.45	06/09/2015	36	...	92	486	16	3	6198
2015360751	1.00	06/09/2015	36	...	404	420	15	1	430
2015360770	5.94	08/09/2015	36	...	87	194	18	3	0
2015360794	3.20	28/09/2015	36	...	89	428	21	4	0
2015360819	2.26	27/12/2015	36	...	70	85	7	0	0

TABLA 1: DATOS INICIALES DE INCENDIOS DE GALICIA

La primera observación es que, entre esos años, en Galicia se han registrado **24.587** incendios de los 82.640 totales de nuestros datos, lo que significa aproximadamente un **30%** del total. En cuanto a términos de superficie, el **22.35%** de la superficie quemada en este data set es gallega.

Antes de comenzar con un análisis más profundo, realizaremos una limpieza de los datos en estudio. En primer lugar, eliminamos los incendios cuyas coordenadas no están definidas, así como los incendios ubicados fuera de la comunidad, ya que en un futuro pueden ser considerados como outliers y pueden proporcionarnos información errónea.

Seleccionamos para nuestro estudio las columnas que consideraremos más relevantes, por este motivo, primeramente, descartaremos el *id* de la comunidad, puesto que no nos va a proporcionar información adicional; posteriormente, prescindiremos de la variable *latlong_explicit* ya que, aunque tengamos ceros en esta variable, consideraremos este punto como un punto del mapa en el que hubo un incendio a pesar de que ese punto no fuera el origen exacto del mismo. Por último, descartaremos también las *causas supuestas* y *causas explícitas*, creando una única variable relevante, denominada *causa descripción*, esta variable se construyó en base a diversos estudios que hacen uso de nuestra fuente de información (Garrido, 2016).

Otra fuente de datos consultada ha sido el IGE (Instituto Galego de Estatística, s.f.), de la cual extraemos información de las distintas superficies totales por municipios, hay que tener en cuenta que estos datos solamente los disponemos para incendios posteriores al año 2005. Debido a que la información viene agrupada por municipio, para poder unirlas a nuestro data set generaremos los mismos códigos utilizados para los municipios de

nuestra tabla de incendios. Tras llevar a cabo un análisis en profundidad, dichos códigos corresponden a la concatenación del *id de provincia* con el *id del municipio* de nuestra tabla original con una pequeña peculiaridad, el código formado posee 5 cifras, por lo que entre los *ids* mencionados se introducen tantos ceros como sea necesario para llegar a obtener la longitud deseada. Bandonos en esta conclusión creamos el "código postal" en la tabla de incendios y, en ese momento, ya somos capaces de unificar ambos data sets.

A continuación, y para enriquecer nuestra base de datos, siguiendo por la geografía gallega hemos decidido añadir datos tanto de embalses, como de los ríos de Galicia. A través del IGN (Instituto Geográfico Nacional, s.f.) obtuvimos los datos correspondientes a los embalses de las provincias de A Coruña, Lugo y Ourense, y, a través de la página web (Embalses, 2021), completamos esta información, añadiendo la información de los embalses de Pontevedra.

Al igual que para los embalses, hemos extraído un conjunto de datos de los ríos gallegos de (Instituto Geográfico Nacional, s.f.) que, haciendo uso de diversas páginas webs como *Wikipedia* o *Galiceando*, se ha completado esta tabla con los diferentes municipios por los que pasa cada río. Este dato ha sido incluido a cada incendio como el *número de ríos* que pasan por su municipio correspondiente.
Por último, ha sido creada una base de datos independiente obtenida de los resúmenes climatológicos mensuales (Meteogalicia Tiempo, s.f.), esta información viene asociada a cada estación meteorológica distribuida en toda la comunidad durante la ventana temporal del que disponemos de incendios, además hemos incluido en este data set las coordenadas correspondientes cada estación (Meteogalicia Estaciones, s.f.).

La asignación tanto embalses como condiciones meteorológicas a cada incendio será determinada por la distancia mínima entre cada embalse o estación y los focos de incendios. Es por este motivo por el que hemos definido una función distancia, basada en la fórmula *Haversine* y una función de distancia mínima para localizar la estación o el embalse más próximo al foco en estudio. Para mejorar el rendimiento de la función de mínima distancia, esta será aplicada siempre entre elementos de la misma provincia.

Análisis descriptivo

Una vez obtenida una base de datos completa y robusta, empezamos a realizar un primer estudio de los datos, podemos observar que los incendios no se distribuyen por igual entre las distintas provincias gallegas. En la **¡Error! No se encuentra el origen de la referencia.**, podemos observar que Ourense destaca en número de incendios ante las demás provincias, curiosamente, la única provincia de interior. Se podría pensar que esto es debido a que Ourense es la provincia que posee mayor superficie forestal, pero en la Figura 2, observamos que Lugo es quien más superficie forestal posee y, a la vez, la que menos incendios sufre en este período de tiempo.

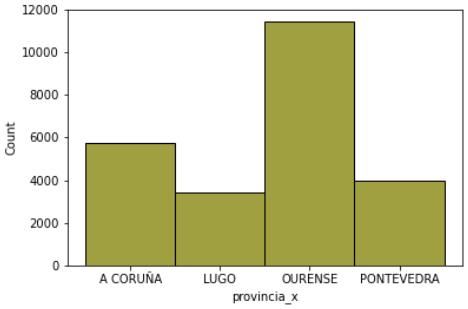


FIGURA 1: HISTOGRAMA
INCENDIOS

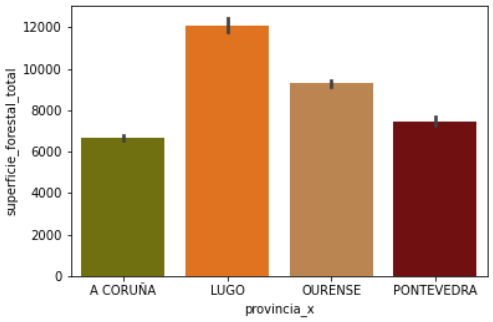


FIGURA 2: DIAGRAMA DE BARRAS
SUPERFICIE FORESTAL

Realizando una vista general en el histórico, observamos en la Figura 3 que el mes en el que se origina el mayor número de incendios es agosto, seguido por septiembre y curiosamente marzo, y en la Figura 4 vemos que la mayor cantidad de incendios suceden en los primeros años de nuestro estudio, entre 2001 y 2006.

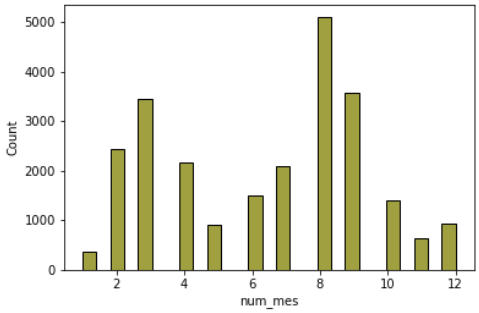


FIGURA 3: HISTOGRAMA
INCENDIOS POR MES

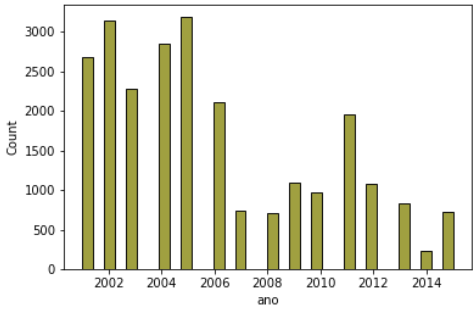


FIGURA 4: HISTOGRAMA
INCENDIOS POR AÑO

En términos de superficie, destaca un incendio ocurrido en el año 2006 en el que han sido quemadas 7000 *ha* que, como podemos observar en el siguiente cuadro, es comparable con la suma de los máximos del resto de provincias; no obstante, en términos del total de hectáreas quemadas sigue destacando Ourense ante las otras:

	superficie					id
	sum	max	min	mean	median	count
provincia_x						
A CORUÑA	78439.18	2842.00	1.0	13.622643	2.5	5758
LUGO	40310.12	2364.67	1.0	11.748796	2.5	3431
OURENSE	165740.40	3236.70	1.0	14.501741	2.5	11429
PONTEVEDRA	82433.96	7316.77	1.0	20.790406	3.0	3965

TABLA 2: RESUMEN DE LOS INCENDIOS AÑO 2006

Al igual que con el número de incendios, en la Figura 5 se puede apreciar que el mes de agosto es el mes que más superficie se ha quemado seguido de octubre. Anualmente, observamos que el año 2006 es el año donde se ha quemado el mayor número de superficie, podemos pensar que esto es debido al incendio mencionado anteriormente, pero analizando las distintas provincias, salvo Ourense todas alcanzan su pico más alto en este año, véase la Figura 6 y Figura 7.

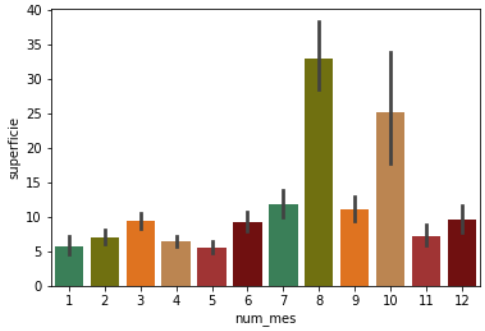


FIGURA 5: DIAGRAMA DE BARRAS
SUPERFICIE POR MES

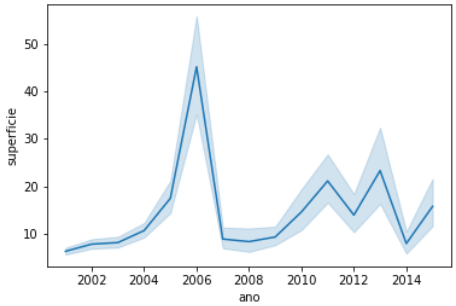


FIGURA 6: GRÁFICO DE LÍNEAS
SUPERFICIE POR AÑO

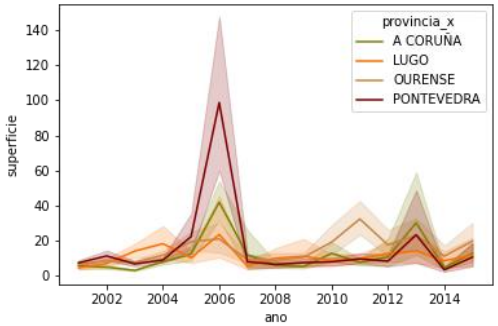


FIGURA 7: GRÁFICO DE LÍNEAS
SUPERFICIE POR AÑO Y PROVINCIA

A continuación, describiremos las principales causas de incendios se dividen en este conjunto de datos en seis categorías, cuya información hemos extraído de (Garrido, 2016):

1. Rayo
2. Descuidos humanos
3. Accidentes
4. Intencionado
5. Desconocido
6. Reproducido

En la Figura 8 podemos observar el dato que posiblemente más llame la atención, y es que más del 85% de los incendios en Galicia durante los años en estudio han sido intencionados. De la Tabla 3 podemos concluir que la mayor parte de los incendios afectan a pequeñas superficies, ya que la mediana de los incendios provocados es de 2.57 *ha*, algo que no difiere mucho de las medianas de las hectáreas calcinadas en cada provincia.

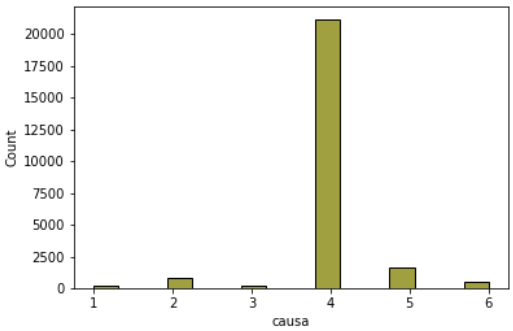


FIGURA 8: HISTOGRAMA
INCENDIOS POR CAUSA

	superficie				id	
	sum	max	min	mean	median	count
causa						
1	5489.40	480.00	1.0	22.314634	3.00	246
2	8279.39	804.94	1.0	10.234104	2.30	809
3	2227.21	245.05	1.0	11.421590	3.00	195
4	308473.28	7316.77	1.0	14.562304	2.57	21183
5	26696.01	1573.00	1.0	16.179400	2.50	1650
6	15758.37	3236.70	1.0	31.516740	2.50	500

TABLA 3: RESUMEN INCENDIOS POR CAUSA

Otro dato relevante en un incendio es su tiempo de control y de extinción, ya que esto, junto a otros factores, puede influir directamente en la superficie quemada. En la Figura 9 y Figura 10 observamos la relación entre estas variables, este último gráfico corrobora el resultado esperado, el tiempo de extinción siempre es superior o igual al de control.

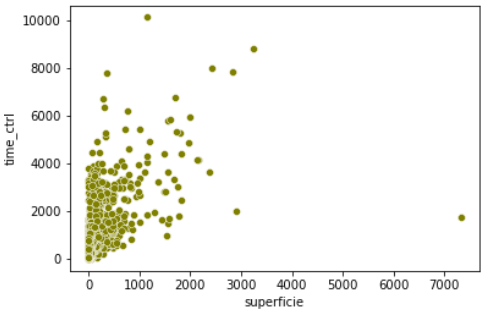


FIGURA 9: GRÁFICO DE DISPERSIÓN
TIEMPO DE CONTROL FRENTE SUPERFICIE AFECTADA

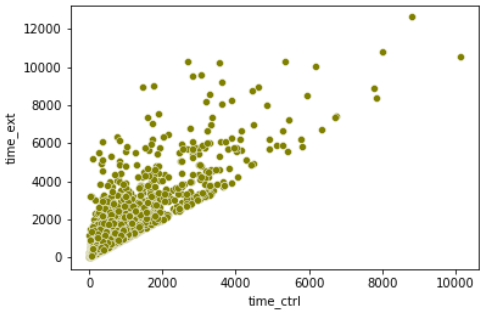


FIGURA 10: GRÁFICO DE DISPERSIÓN
TIEMPO DE CONTROL FRENTE TIEMPO DE EXTINCIÓN

Por otro lado, en el análisis nos hemos percatado de que la presencia de ríos tiene mucha influencia en las demás variables relevantes, pues el número de ríos en los municipios en los que se origina un fuego influye tanto en la superficie quemada, como en tiempos de control y de extinción. En las siguientes gráficas (Figura 14, Figura 12, Figura 13 y Figura 11) observamos la importancia de la presencia de ríos en territorio gallego:

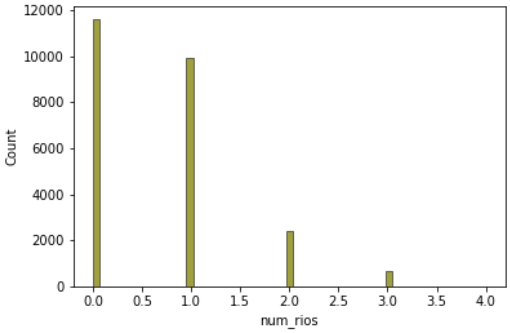


FIGURA 12: HISTOGRAMA
INCENDIOS FRENTE NÚMERO DE RÍOS

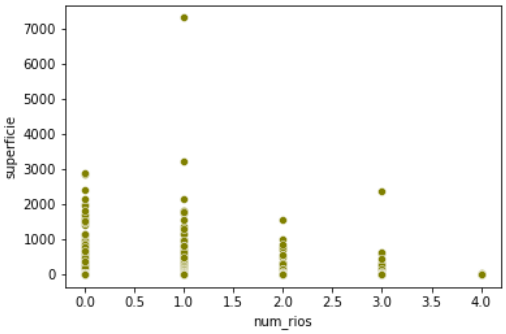


FIGURA 13: DIAGRAMA DE DISPERSIÓN
SUPERFICIE FRENTE NÚMERO DE RÍOS

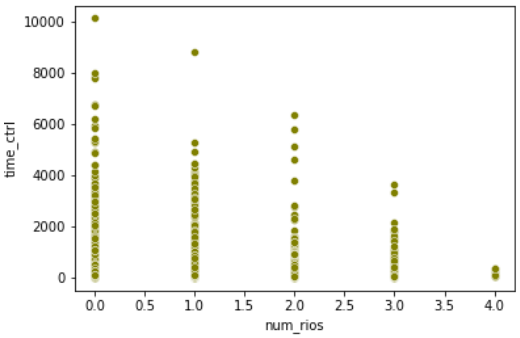


FIGURA 11: DIAGRAMA DE DISPERSIÓN
TIEMPO DE CONTROL FRENTE NÚMERO DE RÍOS

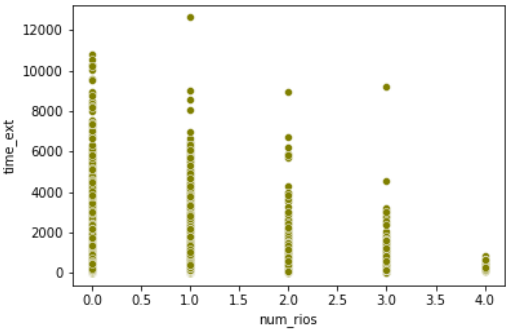


FIGURA 14: DIAGRAMA DE DISPERSIÓN
TIEMPO DE EXTINCIÓN FRENTE NÚMERO DE RÍOS

Por último, introducimos al data set de incendios los datos climáticos del mes en que se produjeron los mismos. Esta información se corresponde con la temperatura, medida en grados centígrados; la precipitación acumulada, en litros por metro cuadrado; el porcentaje de humedad relativa, la velocidad del viento en kilómetros hora, la presión, en hectopascales y, por último, los días acumulados de lluvia y de heladas.

De estos datos vemos que los más relevantes son la humedad máxima y la precipitación acumulada. Podemos observar en la Figura 19, la mayoría de los incendios ocurren cuando la humedad máxima es mayor. Por último, en la Figura 20, vemos que en los meses con menor precipitación acumulada se concentra la mayor cantidad de incendios.

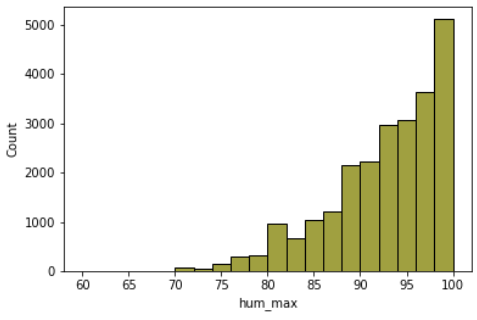


FIGURA 15: HISTOGRAMA
INCENDIOS FRENTE HUMEDAD

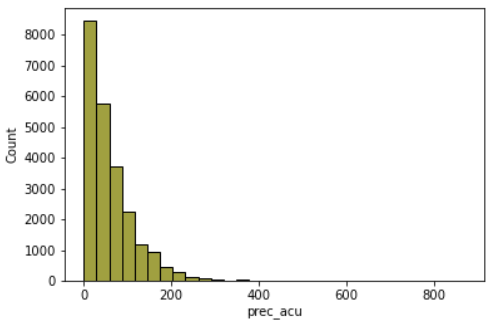


FIGURA 16: HISTOGRAMA
INCENDIOS FRENTE PRECIPITACIÓN

Capítulo 2
Predicción de incendios

A pesar de que el mayor número de incendios ha sido provocado por el ser humano, hemos querido ver la influencia que tienen los datos climáticos a la hora de originarse un fuego. Como hemos comentado en el anterior capítulo, contamos con la base de datos de incendios y la información mensual registrada en las distintas estaciones meteorológicas de Galicia.

Para poder analizar si un incendio se produce en ciertas condiciones medioambientales, es necesario tener no solo datos de los incendios, si no también, tener datos de los **no** incendios. En primer lugar, hemos seleccionado el mínimo número de variables necesario para optimizar el tiempo de ejecución; por lo que crearemos un data frame "*Incendios totales*" con los identificadores de incendio, su fecha (año y mes), el identificador de la provincia y sus coordenadas. Paralelamente guardamos en otro data set las coordenadas únicas de los focos de incendio junto a su identificador de provincia.

Hemos denominado a nuestra variable objetivo como "*lume*" (incendio en gallego) y, puesto que nuestros datos climáticos son mensuales, vamos a crear un data set mensual que contenga todos los incendios ocurridos al mes y los datos climatológicos de la estación más próxima a cada uno de estos focos. Nuestra nueva tabla de datos la creamos mediante un bucle que recorre todos los meses de la ventana temporal dada y en cada mes, primeramente, guarda las variables mencionadas anteriormente junto con la nueva variable *lume=1* en las coordenadas que hubo incendio. Para la variable *lume=0*, hemos seleccionado las coordenadas únicas de todos los focos producidos a lo largo de estos años y elegido aquellas que no están registradas en ese mes en "*Incedios totales*", lo que significaría que aunque es un foco de incendio, dicho mes no fue afectado. Véase el código a continuación:

```
# Incendios totales
inc_totales=df[['id','mes','idprovincia','lat','lng']]
inc_totales['coord']=list(zip(inc_totales.lat, inc_totales.lng))
inc_totales=inc_totales.reset_index()

#Localización única de incendios
localizacion_inc_unic=inc_totales[['coord','lat','lng','idprovincia']].
drop_duplicates().reset_index()
```

```
    inc=pd.DataFrame()
    for i in inc_totales['mes'].unique():
        print('Mes: '+str(i))
        #Inclusión de los incendios del mes i
        inc_i=inc_totales[inc_totales['mes']==i][['mes','idprovincia','lat',
    'lng','coord']]
        inc_i['lume']=1
        inc=inc.append(inc_i)

        #Inclusión de los NO incendios del mes i
        index_noinc=np.vectorize(lambda t: t not in inc_totales[inc_totales
```

```
    ['mes']==i]['coord'].tolist())
12      noinc_i=localizacion_inc_unic[index_noinc(localizacion_inc_unic['coord'])]
13      noinc_i['mes']=i
14      noinc_i['lume']=0
15      inc=inc.append(noinc_i[['mes','idprovincia','lat','lng','coord','lume']])
```

Una vez obtenido el data set con la variable objetivo, vamos a asignar a cada registro los datos temporales en ese mes. Para ello hemos aplicado la función del semiverseno, también conocida como función *haversine*, que calcula la distancia entre dos puntos de la superficie terrestre:

```
1 from math import radians, cos, sin, asin, sqrt
2 def distance(lat1, lat2, lon1, lon2):
3
4     # convertimos los grados en radianes.
5     lon1 = radians(lon1)
6     lon2 = radians(lon2)
7     lat1 = radians(lat1)
8     lat2 = radians(lat2)
9
10    # Haversine formula
11    dlon = lon2 - lon1
12    dlat = lat2 - lat1
13    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
14
15    c = 2 * asin(sqrt(a))
16
17    # Radio terrestre
18    r = 6371
19
20    # calculate the result
21    return(c * r)
```

Inicialmente, uniremos la tabla de incendios con las condiciones meteorológicas por medio de la provincia y del mes correspondiente al incendio y a los datos climatológicos de cada estación. Una vez tengamos la tabla conjunta, se aplicará la función de distancia a todas las coordenadas de los distintos registros, calculando así la distancia entre cada incendio y todas las estaciones localizadas en la misma provincia. Finalmente, aplicaremos la función de distancia mínima que mostramos a continuación para que a cada foco de incendio le corresponda unas condiciones climáticas únicas:

```
1  def dist_min(df1,column_lat1,column_lat2,column_lng1,column_lng2,index,name):
2
3      #Add column containing distances
4      df1[name]=df1.apply(lambda x: distance(x[column_lat1], x[column_lat2],
                                  x[column_lng1], x[column_lng2]), axis=1)
5      #Selecting the min distance per id
6      df1_min = df1.groupby([index]).agg({name: 'min'})
7      df2 = pd.merge(df1, df1_min, on = index, how ='inner')
8      df = df2[(df2[name+'_x']==df2[name+'_y']) | df2[name+'_x'].isna()]
9      return(df)
```

Una vez hecho esto, obtenemos la tabla final que consta de 2.046.302 filas, de las cuales hay 24.583 incendios (*lume=1*), y las siguientes columnas:

- Temperatura media
- Temperatura máxima media
- Temperatura mínima media
- Precipitación acumulada
- Humedad media
- Humedad máxima
- Humedad mínima
- Velocidad del viento media
- Presión
- Lluvia
- Helada
- Lume (variable objetivo)

Inicialmente usaremos el método de clasificación conocido como regresión logística, modelo en el que se predice una respuesta binaria (nuestra variable *lume*) en función de las demás variables independientes, la regresión logística está incluida entre los modelos lineales generalizados y se utiliza en gran medida para el modelado de la probabilidad de que suceda, en nuestro caso, un incendio bajo unas condiciones determinadas (Cramer, 2002).

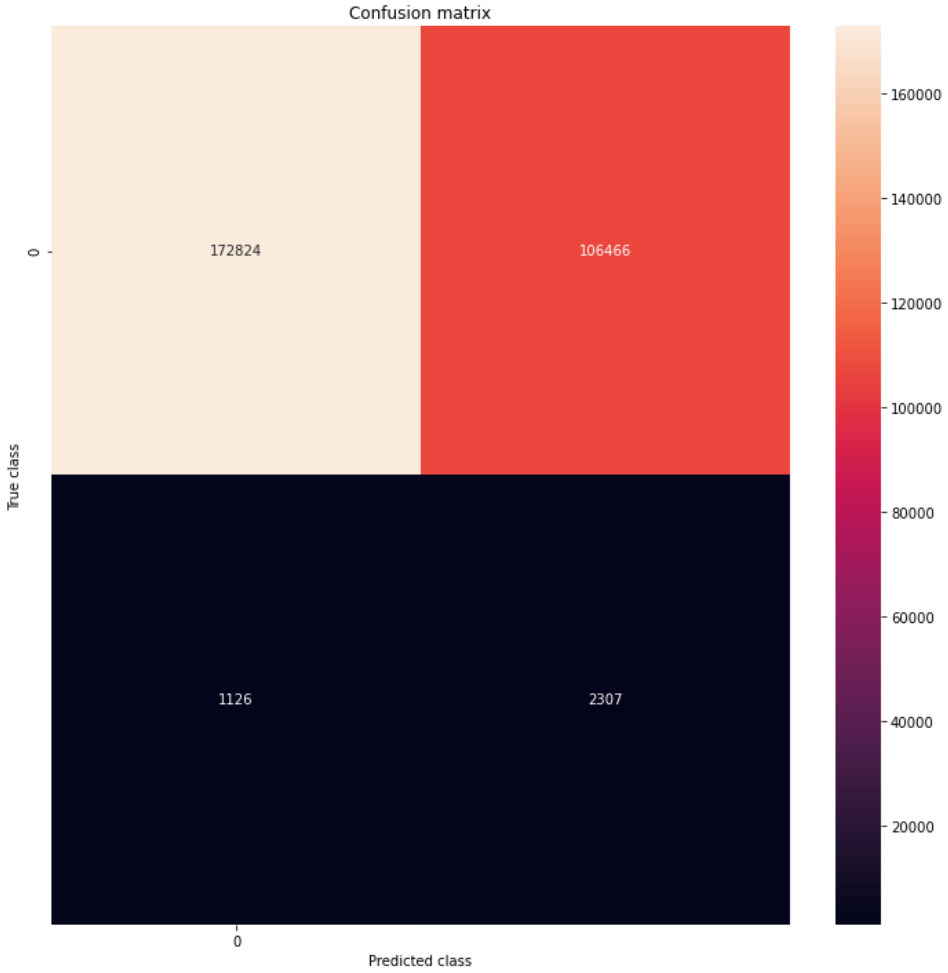
Regresión Logística

Para usar el método *LogisticRegression* de *Sklearn*, en primer lugar hemos eliminado los registros que tuvieran algún valor NaN, quedándonos así con un conjunto de 14.1611 registros. Hemos dividido el conjunto total de datos en un primer set de entrenamiento y otro de test con una proporción de 80% a 20%, respectivamente, y obtuvimos los siguientes resultados:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	279368
1	0.00	0.00	0.00	3355
accuracy			0.99	282723
macro avg	0.49	0.50	0.50	282723
weighted avg	0.98	0.99	0.98	282723

Esto es, debido a la gran diferencia entre incendios y no incendios, el método ha asignado a todas las clases del conjunto de test un 0.

Dados estos resultados, se ha obtado por aplicar el mismo modelo, pero balanceando las clases minoritarias con *RandomOverSampler* para ver cómo podría mejorar así el modelo:



	precision	recall	f1-score	support
0	0.99	0.62	0.76	279290
1	0.02	0.67	0.04	3433
accuracy			0.62	282723
macro avg	0.51	0.65	0.40	282723
weighted avg	0.98	0.62	0.75	282723

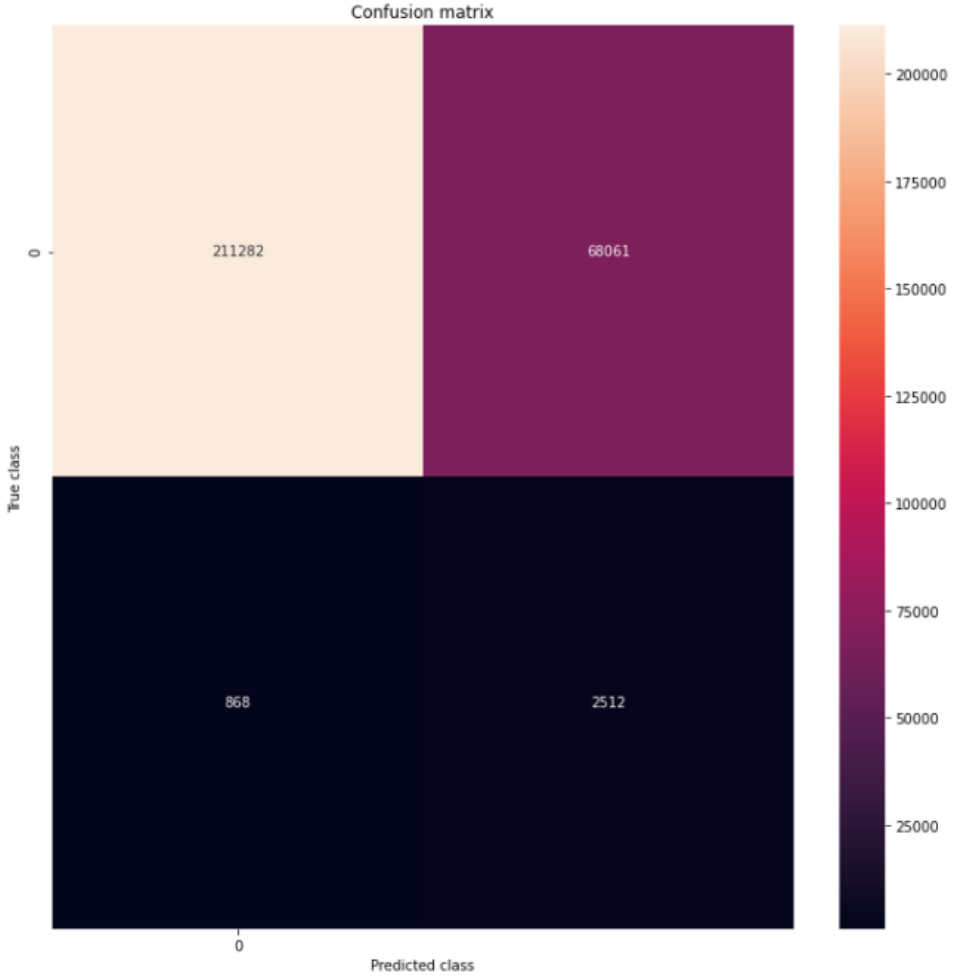
Podemos observar cómo, balanceando las clases, mejora el modelo, obteniendo además un valor de recall del 67%, es decir, que el modelo predice el 67% de los incendios reales. En nuestro caso, es más importante maximizar el recall, antes que la precisión de la categoría 1, ya que consideraremos más importante detectar el máximo número de incendios reales que minimizar los falsos positivos.

XGBoost – Sin NaN

Intentando obtener un modelo óptimo para nuestros datos, aplicamos el método *XGBClassifier* de *XGBoost*, para regresión logística. A esto, para obtener los parámetros más adecuados le vamos a aplicar validación cruzada con *GridSearchCV* con “recall” como scoring. Una vez entrenado el modelo, obtenemos los mismos resultados que en

nuestro primer intento. Lo que vemos es que, sin balancear las clases, éstos dos métodos se comportan de la misma forma.

Repetimos el proceso aplicando de nuevo *RandomOverSampling* sobre la clase minoritaria y vemos como obtenemos una mejora en nuestros resultados:



	precision	recall	f1-score	support
0	1.00	0.76	0.86	279343
1	0.04	0.74	0.07	3380
accuracy			0.76	282723
macro avg	0.52	0.75	0.46	282723
weighted avg	0.98	0.76	0.85	282723

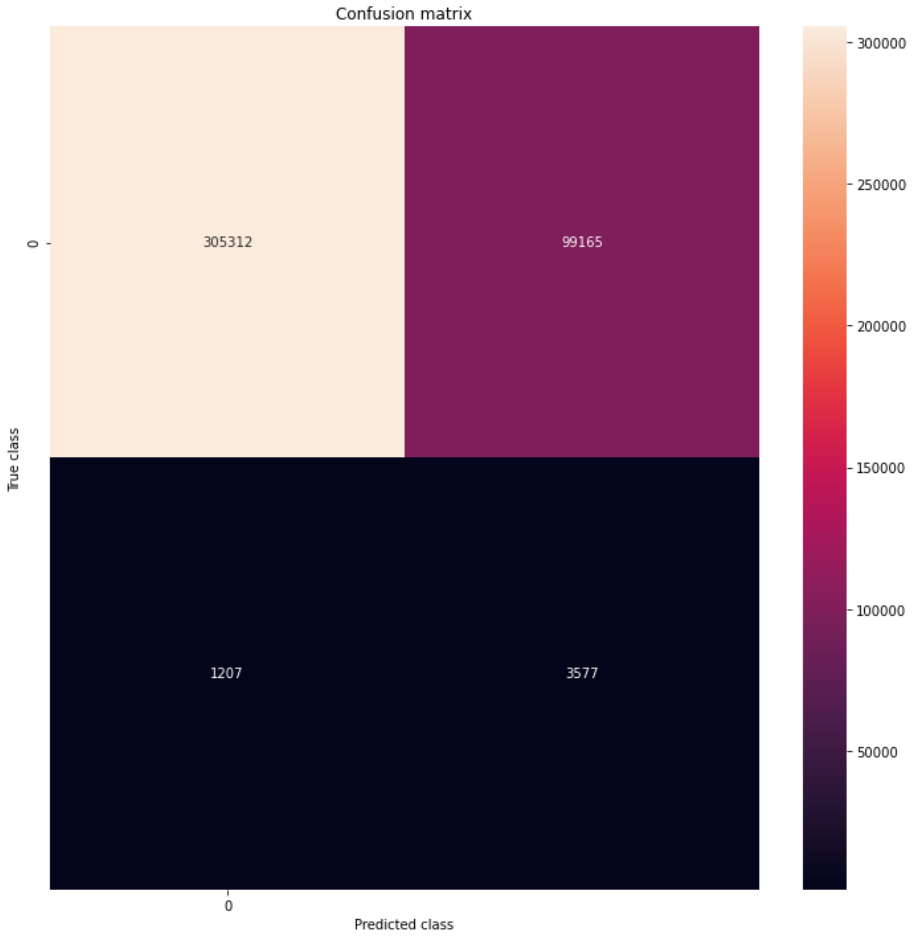
Como se pudo observar, hemos duplicado nuestra precisión y el recall aumenta al 74%.

No obstante, todo este estudio lo hemos realizado eliminando todos los registros que contengan algún valor nulo; pero puesto que el modelo *XGBClassifier* es capaz de identificar tendencias dentro de estos valores ausentes, hemos decidido replicar el proceso con todos los registros de la tabla.

XGBoost – Con NaN

Replicando el modelo con los datos sin balancear, obtenemos los mismos resultados que anteriormente, este resultado, con cada intento de mejora que vamos aplicando durante nuestro estudio, se mantiene siempre inmóvil, por lo que a partir de ahora omitiremos continuar hablando de datos sin balancear.

Repitiendo el procedimiento anterior, con las mismas condiciones, pero esta vez, incluyendo los datos nulos, para ver cómo responde nuestro modelo. Logramos subir un 1% el recall aunque bajamos al 3% la precisión, lo cual es una leve mejora:



	precision	recall	f1-score	support
0	1.00	0.75	0.86	404477
1	0.03	0.75	0.07	4784
accuracy			0.75	409261
macro avg	0.52	0.75	0.46	409261
weighted avg	0.98	0.75	0.85	409261

Analizando los resultados en profundidad, observamos la distribución de los verdaderos y los falsos positivos. Apreciamos que a pesar de tener muy poca precisión, el número de verdaderos positivos que han sido detectados entre un 50% y un 60% de probabilidad de incendio es muy elevado (568) como podemos observar en la siguiente tabla:

	50-60		60-70		70-80		80-90		90-100	
	sum	prop	sum	prop	sum	prop	sum	prop	sum	prop
real										
1	568	15.879228	740	20.687727	1011	28.263908	1070	29.913335	188	5.255801

Y puesto que tenemos un 25% de incendios reales que no podemos detectar, y en nuestro caso, necesitamos minimizar los falsos negativos, ya que consideramos que los falsos positivos los interpretaremos como una alerta a ese foco en un determinado momento.

Para mejorar el recall, buscaremos una opción para calcular la probabilidad optima de corte. Es por ello por lo que vamos a crear una función para obtener la precisión, recall y el f1-score en función de la probabilidad de corte, además este programa optimiza el modelo según argumento que se quiera maximizar y devuelve la probabilidad de corte optima:

```
def predict_custom(modelo,X,prob=0.5):
    y_prob = modelo.predict_proba(X)[:,1]
    y_pred = np.where(y_prob>=prob,1,0)
    return y_pred
```

```
def arg_model(y_pred,y_real):
    resul = pd.DataFrame()
    resul['prediction'] = y_pred
    resul['real'] = np.array(y_real)

    #Cálculo de FalsePositive, TruePositive, FalseNegative, TrueNegative
    FP = resul[(resul['prediction']!=resul['real'])&(resul['prediction']==1)].shape[0]
    TP = resul[(resul['prediction']==resul['real'])&(resul['prediction']==1)].shape[0]
    FN = resul[(resul['prediction']!=resul['real'])&(resul['prediction']==0)].shape[0]
    TN = resul[(resul['prediction']==resul['real'])&(resul['prediction']==0)].shape[0]

    #Cálculo de los argumentos
    precision = TP / (TP+FP)
    recall = TP / (TP+FN)
    f1 = 2 * recall * precision / (recall + precision)
    return precision,recall,f1

```

Observamos al comparar la probabilidad de corte entre 45%, 50% y 55% que el valor con mejor recall, y sin penalizar demasiado la precisión es el 45%, por lo que consideraremos el corte en 45% obteniendo así un recall del 80%.

```
i = 0.45
y_pred = predict_custom(modelo=clf_f1, X=X_test, prob=i)
precision,recall,f1 = arg_model(y_pred,y_test)

print('Precisión: '+str(precision)+', Recall: '+str(recall)+', f1:
  '+str(f1))))
```

```
Precisión: 0.030264438098703 52, Recall: 0.7997491638795987,
f1: 0.058321837153113884
```

```
i = 0.5
y_pred = predict_custom(modelo=clf_f1, X=X_test, prob=i)
precision,recall,f1 = arg_model(y_pred,y_test)

print('Precisión: '+str(precision)+', Recall: '+str(recall)+', f1:
  '+str(f1))
```

```
Precisión: 0.034815362753304394, Recall: 0.747700668896321,
f1: 0.06653274556851366
```

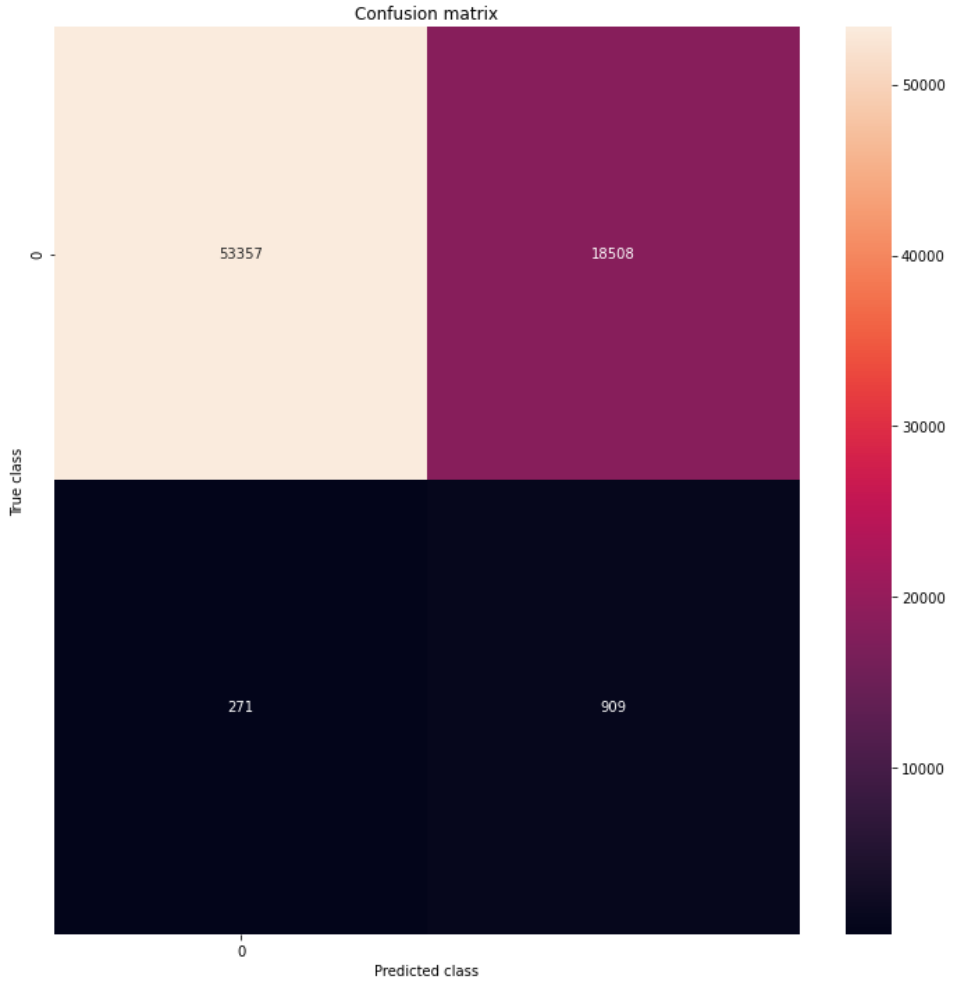
```
i = 0.55
y_pred = predict_custom(modelo=clf_f1, X=X_test, prob=i)
precision,recall,f1 = arg_model(y_pred,y_test)
print('Precisión: '+str(precision)+', Recall: '+str(recall)+', f1:
  '+str(f1))
```

```
Precisión: 0.039371575932594154, Recall: 0.6925167224080268, f1: 0.
07450720221295162
```

XGBoost – Provincias

Una vez estudiado esto, nos planteamos ahora si el modelo predecirá mejor los incendios si tratamos a cada una de las cuatro provincias por separado, ya que no todas sufren el mismo número de incendios y las condiciones meteorológicas difieren considerablemente entre unas y otras. Repetimos el proceso entonces para A Coruña, Lugo, Ourense y Pontevedra obteniendo los siguientes resultados:

- En A Coruña, los resultados para la probabilidad de corte del 50% serían una precisión del 5% y un recall del 77%, que también es ligeramente superior a los resultados obtenidos considerando toda la comunidad.

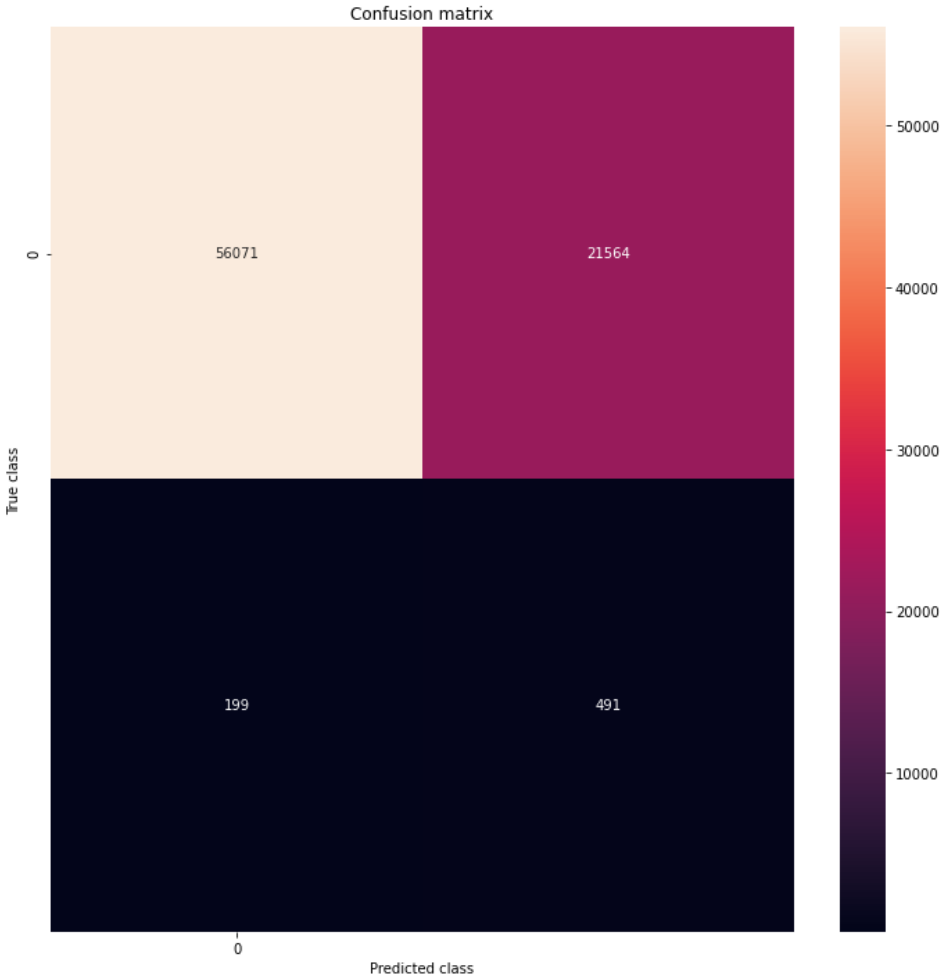


	precision	recall	f1-score	support
0	0.99	0.74	0.85	71865
1	0.05	0.77	0.09	1180
accuracy			0.74	73045
macro avg	0.52	0.76	0.47	73045
weighted avg	0.98	0.74	0.84	73045

Por lo que hemos comentado anteriormente, también vamos a calculas dichas métricas para el 45% de probabilidad de corte, en este caso obtenemos una precisión del 4% y mantenemos el recall en un 80%, por lo que de momento nuestro modelo mejora ligeramente ya que aumentamos la precisión en un 1%.

Precisión: 0.04232066561727007, Recall: 0.7974576271186441, f1: 0.08037582746102925

- En Lugo, que es la provincia con menos incendios y menos hectáreas quemadas, hemos obtenido como resultados, 71% de recall y un 2% de precisión. Lo que nos hace pensar que puede que no sea ésta la forma óptima para nuestro análisis.

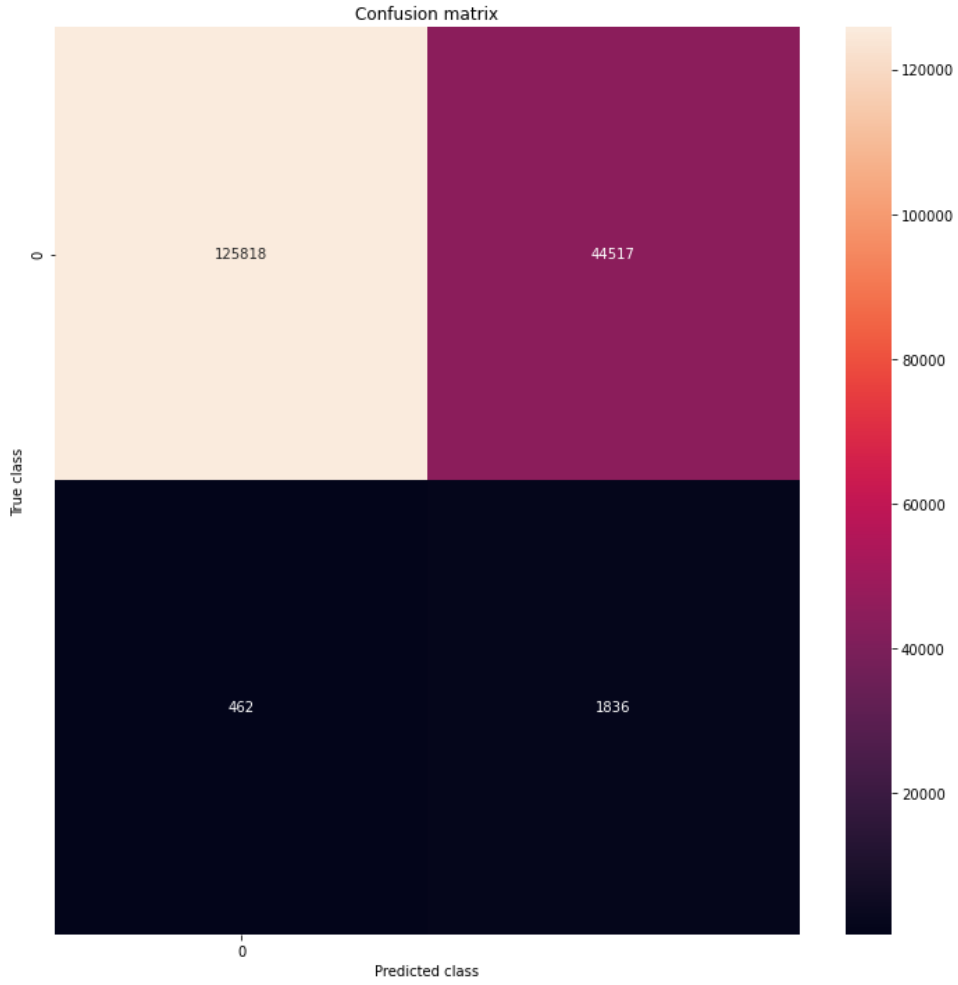


	precision	recall	f1-score	support
0	1.00	0.72	0.84	77635
1	0.02	0.71	0.04	690
accuracy			0.72	78325
macro avg	0.51	0.72	0.44	78325
weighted avg	0.99	0.72	0.83	78325

Analizando los resultados para el porcentaje de corte del 45% mantenemos el 2% de precisión y logramos aumentar el recall hasta el 75%, también por debajo de los valores de las métricas de toda la comunidad:

Precisión: 0.02082156403841498, Recall: 0.7478260869565218, f1: 0.040515075376884424

- Continuamos con Ourense, la provincia con mayor número de incendios. En ella, la precisión y recall obtenidos con la probabilidad del 50% son 4% y 80% respectivamente.

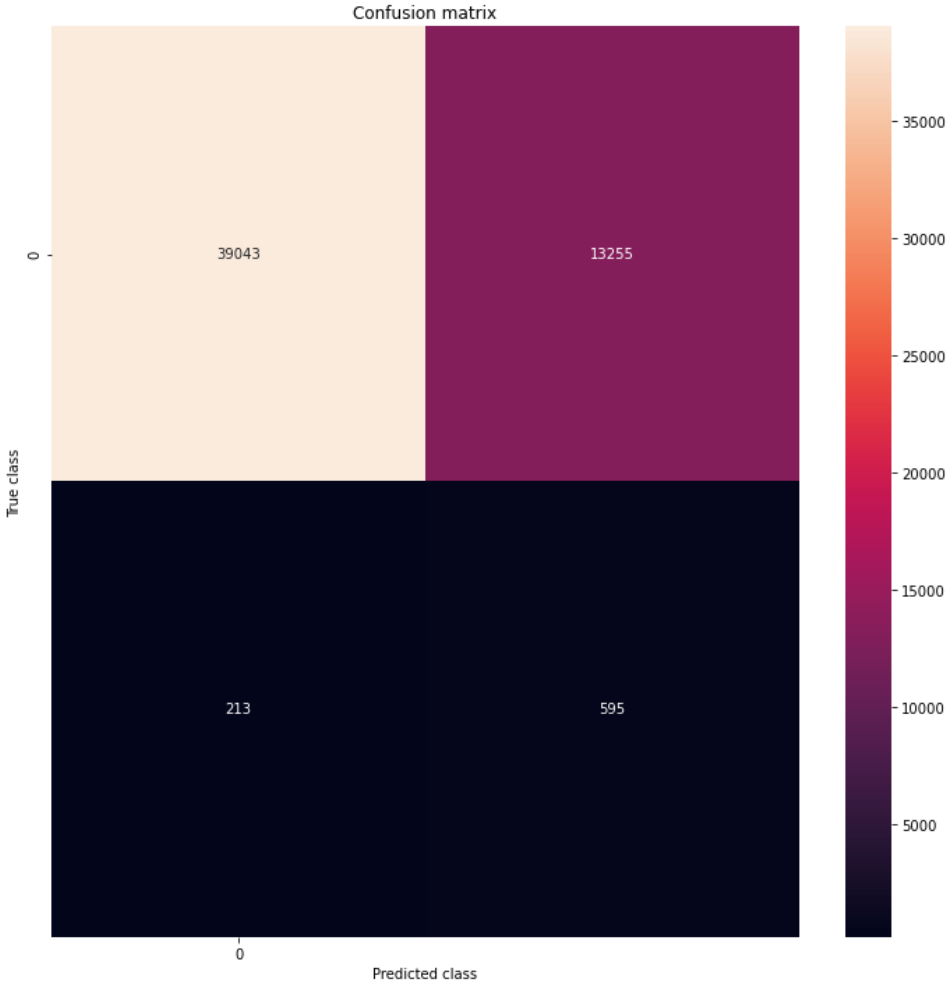


	precision	recall	f1-score	support
0	1.00	0.74	0.85	170335
1	0.04	0.80	0.08	2298
accuracy			0.74	172633
macro avg	0.52	0.77	0.46	172633
weighted avg	0.98	0.74	0.84	172633

En cambio, al aplicar el corte en 45%, mantenemos el 4% de precisión, pero el recall sube a un 83%. Vemos como mejora el modelo donde mayor cantidad de incendios se concentran:

Precisión: 0.037050016548877594, Recall: 0.8281114012184508, f1: 0.07092674381767018

- Por último, en Pontevedra obtenemos los siguientes datos:



	precision	recall	f1-score	support
0	0.99	0.75	0.85	52298
1	0.04	0.74	0.08	808
accuracy			0.75	53106
macro avg	0.52	0.74	0.47	53106
weighted avg	0.98	0.75	0.84	53106

Para un 45% de probabilidad de corte obtenemos un 4% de precisión y un 79% de recall:

Precisión: 0.039175766549447835, Recall: 0.7858910891089109, f1: 0.07463125110183932

Comparativa entre modelos

A continuación, presentamos un cuadro comparativo que contiene las métricas calculadas en cada modelo construido:

Modelo	**Precisión**	**Recall**	**F1-score**
Regresión logística	0.02	0.67	0.04
XGBoost – Sin NaN	0.04	0.74	0.07
XGBoost – Con NaN	0.03	0.75	0.07
XGBoost – Provincias – A Coruña	0.05	0.77	0.09
XGBoost – Provincias – Lugo	0.02	0.71	0.04
XGBoost – Provincias – Ourense	0.04	0.80	0.08
XGBoost – Provincias – Pontevedra	0.04	0.74	0.08

Observando el cuadro podemos asegurar entonces que, aunque nuestro modelo XGBoost diferenciando provincias no trata a todas las provincias por igual, es el que mejor se adapta a nuestro modelo y por lo tanto es el aplicaremos en la página web que construiremos en el capítulo 4.

El motivo por el que decidimos quedarnos con este modelo es porque, aunque empeora el recall de las provincias de Lugo y Pontevedra en relación con el modelo XGBoost – Con NaN, este valor mejora significativamente para las provincias de A Coruña y Ourense, y ambas regiones se corresponden con las más afectadas según el número de incendios producidos.

En un intento de mejora del modelo, hemos probado a incluir el número de ríos en nuestro modelo, ya que hemos visto en el capítulo anterior, que hay una gran influencia en función del número de ríos presentes en el municipio incendiado. Sin embargo, este dato no nos ha aportado ninguna información adicional, además de haber penalizado ligeramente tanto el recall como la precisión en las cuatro provincias gallegas.

Capítulo 3

Geolocalización de torres de control

Actualmente existe un plan de prevención y defensa contra los incendios forestales de Galicia, esta normativa recibe el nombre de PLADIGA y su objetivo es establecer la organización y el procedimiento de actuación de los recursos y servicios para la defensa de los territorios forestales. Entre otras regulaciones, este documento contiene la cantidad de torres de vigilancia fijas que posee Galicia y sus ubicaciones correspondientes (PLADIGA, 2020)

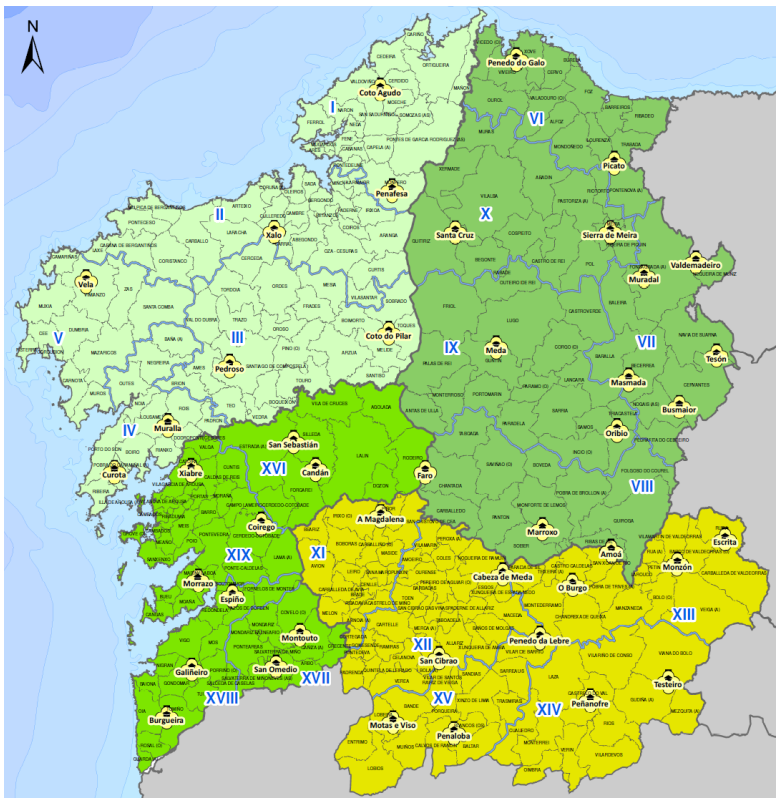


FIGURA 17: UBICACIÓN DE TORRES DE CONTROL

En el presente capítulo se pretende detallar el método utilizado para optimizar la ubicación de un número determinado de torres de control. En particular nos hemos basado en uno de los métodos más utilizados en minería de datos: el método de clasificación K-means, éste busca la mejor partición del conjunto de objetos para un número dado de clases *k*, específicamente encuentra la clasificación que minimiza la suma de cuadrados de las distancias de cada observación al centro de la clase a la que es asignado (Hartigan, 1975).

El algoritmo comienza estableciendo centros iniciales para cada clase elegidos al azar, a continuación, se asigna cada observación a la clase cuyo centro esté más próximo, y finalizada esta asignación se recalculan los centros de cada clase como la media de todos los elementos que la forman. Estos pasos (asignación de cada elemento a una clase y recálculo de centros) se iteran hasta que la clasificación obtenida sea estable. Este

algoritmo reduce en cada iteración la varianza interna de las clases, logrando así que los elementos de cada clase sean similares entre sí, formando clases homogéneas.

Hemos creado una función para reproducir el algoritmo adaptado a la distancia geodésica, dicha distancia ya se ha mencionado en el capítulo 2 donde se creó una función particular para su cálculo basado en la fórmula de *Haversine*.

Además de la subfunción correspondiente a la distancia, se crean las siguientes:

1. Subfunción de asignamiento: con dos argumentos de entrada correspondientes a los centros de las k clases y al data set con las localizaciones de la totalidad de incendios. En este proceso se recorre todas las observaciones asignándoles una clase determinada por la distancia geodésica mínima entre cada incendio y los centros de las k clases.

```
# Función asignar casos a cada centroide
def assignment(df, centroids):
    df_assig=df
    for i in centroids.keys():
        df_assig['distance_from_{}'.format(i)] = df_assig.apply( lambda x:
distance(x['lat'],centroids[i][0],x['lng'],centroids[i][1]), axis=1)

    centroid_distance_cols = ['distance_from_{}'.format(i) for i in centroids.keys()]
    df_assig['closest'] = df_assig.loc[:, centroid_distance_cols].idxmin(axis=1)
    df_assig['closest'] = df_assig['closest'].map(lambda x: int(x.lstrip('distance_from_')))
    return df_assig
```

2. Subfunción de actualización de centros: con dos argumentos de entrada correspondientes a los centros de las k clases y al data set con las localizaciones y las asignaciones de la totalidad de incendios.

```
# Función actualizar los centroides
def update(df_up,centroids):
    for i in centroids.keys():
        centroids[i][0] = np.mean(df_up[df_up['closest'] == i]['lat'])
        centroids[i][1] = np.mean(df_up[df_up['closest'] == i]['lng'])
    return centroids
```

Una vez construidas las subfunciones mencionadas, se crea la función principal simulando el algoritmo de K-means con la distancia geodésica. A continuación, mostramos el código utilizado para su creación:

```

def KMeans_geo(df,k,iter_max=500,cinit_meth='sample_point'):
    df_KMeans_geo=df
    # centroids[i] = [lat, lng]
    if cinit_meth == 'rectangular':
        lat_min= df_KMeans_geo['lat'].min()
        lat_max= df_KMeans_geo['lat'].max()
        lng_min= df_KMeans_geo['lng'].min()
        lng_max= df_KMeans_geo['lng'].max()
        centroids = {

```

```
            i+1: [lat_min+(lat_max-lat_min)*random.random(), lng_min+(lng_max-
lng_min)*random.random()]
            for i in range(k)
        }
    else:
        x = {i+1: np.random.randint(0, len(df_KMeans_geo)-1) for i in range(k)}
        centroids = {
            j+1: [df_KMeans_geo['lat'][df_KMeans_geo.index[x[j+1]]],
df_KMeans_geo['lng'][df_KMeans_geo.index[x[j+1]]]]
            for j in range(k)
        }
    df_KMeans_geo = assignment(df_KMeans_geo, centroids)
    j = 1
    while True:
        centroids_old = df_KMeans_geo['closest'].copy(deep=True)
        centroids = update(df_KMeans_geo,centroids)
        df_KMeans_geo = assignment(df_KMeans_geo, centroids)
        centroids_new = df_KMeans_geo['closest'].copy(deep=True)
        j+=1
        if centroids_old.equals(centroids_new):
            break
        if iter_max == j:
            break
    return(df_KMeans_geo,centroids,j)
```

Como se puede observar en el código y como se adelantó en la descripción del método K-means, las primeras líneas se corresponden a la selección inicial de los centros de las k clases deseadas. Esta elección puede ser por la selección aleatoria de los puntos pertenecientes al rectángulo que contiene todo el territorio gallego afectado y la segunda, es la selección aleatoria de los focos de incendios producidos en el intervalo de tiempo en estudio. A continuación, se utiliza la subfunción de asignación para obtener la primera clasificación de los incendios y tras este primer paso, se utiliza la segunda subfunción creada para recalcular los centroides como la media de las localizaciones de todos los incendios de cada una de las clases. Estos pasos se repiten hasta que se cumplan una de las dos condiciones de parada, que los centroides no cambien o que se llegue hasta la iteración máxima establecida.

Hacemos uso de la función principal para calcular la localización óptima de k torres de vigilancia, donde k recorre el intervalo de [40,50], para así poder valorar la nueva ubicación de otra torre de control, la ubicación optima de las 44 ya existentes o incluso la mejor localización de dichas torres de control en el caso de poseer menos recursos.

A continuación, presentamos las figuras [Figura 18. 40 TorresFigura 18-Figura 28Figura 28] con las localizaciones óptimas obtenidas con la adaptación del método de K-means y además hemos mostrado el radio de control en función de la localización del incendio más lejano de cada clase:

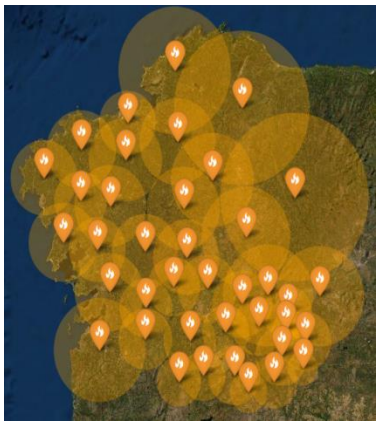


FIGURA 18. 40 TORRES

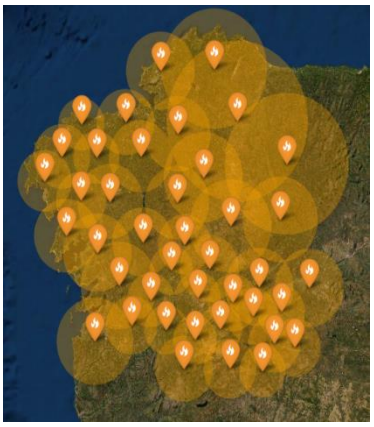


FIGURA 19. 41 TORRES

FIGURA 20. 42 TORRES

FIGURA 21. 43 TORRES

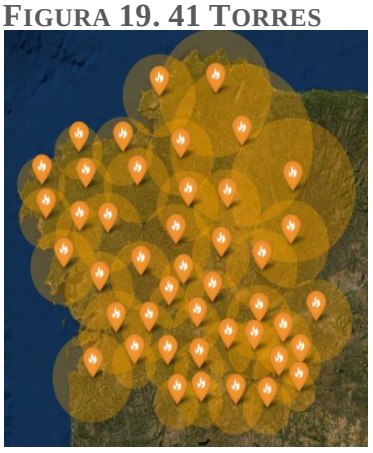


FIGURA 22. 44 TORRES

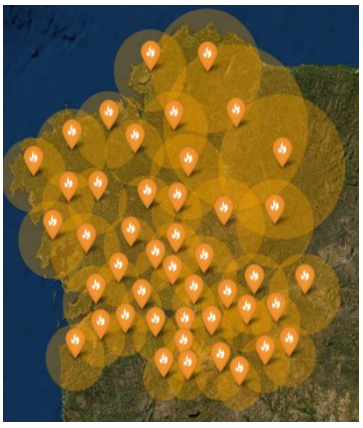


FIGURA 23. 45 TORRES

FIGURA 24. 46 TORRES

FIGURA 25. 47 TORRES

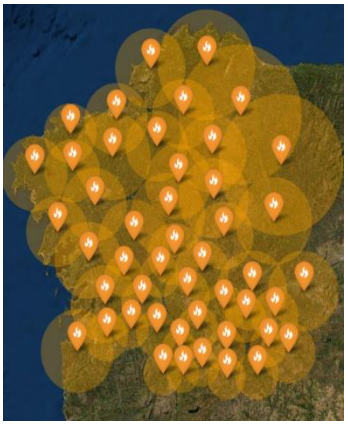


FIGURA 26. 48 TORRES

FIGURA 27. 49 TORRES

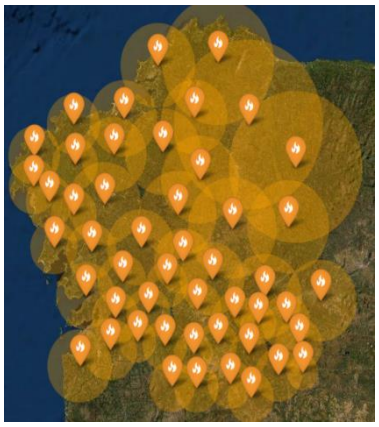


FIGURA 28. 50 TORRES

En todas estas imágenes podemos apreciar la fuerte condensación de las torres de control en la zona inferior del territorio gallego, esta zona se corresponde con Ourense, dicha provincia tiene el papel protagonista de los incendios sufridos estos últimos años.

Para ilustrar la clasificación obtenida hemos creado el gráfico correspondiente a la **¡Error! No se encuentra el origen de la referencia.**, en él representamos los incendios asignados a una torre determinada, además hemos optado por ampliar la zona de la torre elegida para poder enseñar dos evidencias:

1. El foco de incendio más alejado es el que determina el radio de control de la torre.
2. Se puede observar claramente los hiperplanos de separación entre la clase seleccionada y todas las clases que la rodean.

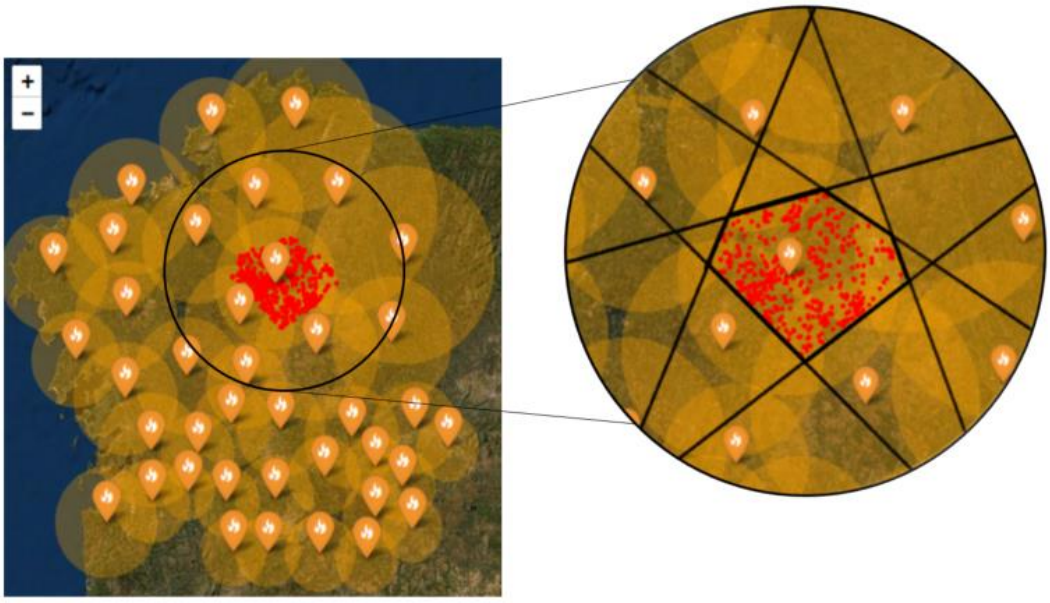


FIGURA 29: INCENDIOS CONTROLADOS POR UNA TORRE ESPECIFICA

Capítulo 4
Front End

Una parte muy relevante en cualquier estudio estadístico es la comunicación correcta de los resultados obtenidos y la capacidad de hacernos entender a cualquier tipo de público tenga o no formación relacionada con la ciencia de datos. Por este motivo, tras realizar los diversos análisis comentados en los capítulos 1, 2 y 3, hemos creado una página web donde los usuarios podrán observar las principales características de los datos y utilizar los diversos métodos creados a lo largo de este trabajo.

Antes de explicar detalladamente las 3 pestañas establecidas, es necesario mencionar que para dicha creación hemos utilizado la librería Streamlit que nos proporciona las herramientas necesarias para desarrollar una página web completa.

Como encabezado hemos hecho primero una pequeña introducción del tema a tratar, posteriormente encontramos una sección oculta con información de interés, en particular esta sección contiene las librerías empleadas a lo largo de este proyecto y las fuentes de datos utilizada para los análisis llevados a cabo, estas fuentes de datos se muestran como hiperlink para redirigir al usuario a la página en cuestión. Por último, se encuentra un desplegable para poder seleccionar la pestaña en la que el usuario desea navegar (ver Figura 30).

FIGURA 30: CABECERO DEL FRONT END

1ª Pestaña: Datos

La primera pestaña se corresponde con un análisis inicial de los incendios ocurridos en un determinado año y en unas determinadas provincias:

FIGURA 31: PESTAÑA DATOS I

En la imagen anterior se puede observar una parte de la primera pestaña, en el lateral izquierdo el usuario podrá seleccionar el año que desea visualizar y las provincias que quiere considerar y a la derecha obtendrá los resultados de dicha elección. Primeramente, se muestra la descripción estadística de las variables cuantitativas que poseemos, en segundo lugar, se muestra un gráfico que contiene el número de incendios anual y mensual de las provincias seleccionadas.

Por ejemplo, los valores establecidos por defecto son los incendios del año 2015 correspondientes a todas las provincias, en la descripción estadística podemos ver la superficie media afectada en dicho año (15.7533 ha) además de otros datos relevantes de esta variable y de otras como puede ser el tiempo de control y el tiempo de extinción. Por otro lado, en el gráfico que se encuentra a la derecha se observa que en ese año Ourense es la provincia más afectada, en particular en el mes 8 (agosto) el número de incendios difiere significativamente con respecto a las demás provincias.

En la misma pestaña el usuario posee un análisis enfocado en la superficie quemada del año y de las provincias seleccionadas. En la Figura 32 se puede observar un análisis descriptivo dividido por provincias para poder comparas entre sí, además de obtener un gráfico para la comparación visual entre provincias.

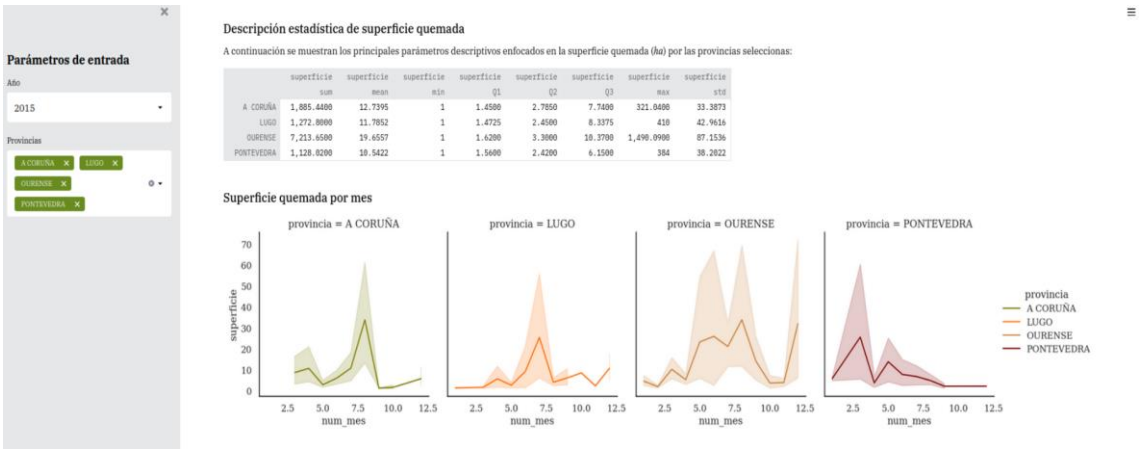


FIGURA 32: PESTAÑA DATOS II

Para mejorar la visualización de esta sección se restructura la página según el número de provincias seccionadas, así, como se aprecia en la Figura 33, si el usuario desea considerar solo las provincias de A Coruña y Lugo, los gráficos se encontrarán a la derecha de la página optimizando tanto el espacio utilizado como la visualización de dichos grafos.

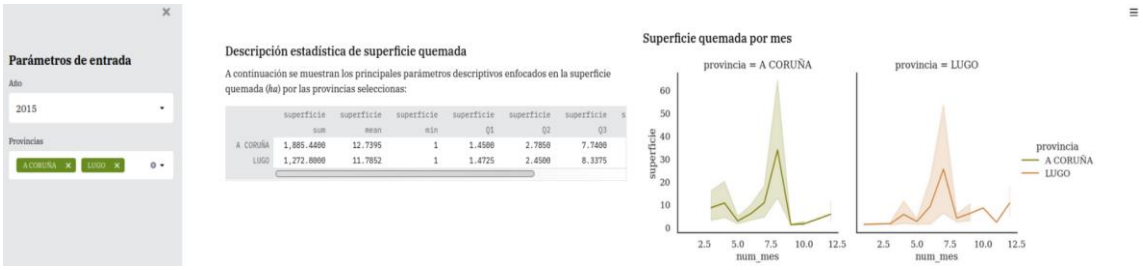


FIGURA 33: PESTAÑA DATOS III

A continuación, se presenta todo el conjunto de datos de los incendios ocurridos en la selección del usuario, indicando las dimensiones de este set. En la Figura 34**¡Error! No se encuentra el origen de la referencia.** se observan los datos de los incendios ocurridos en el año 2015 en las provincias de A Coruña y Lugo, con un total de 256 incendios (filas) y 41 variables (columnas). Además, se ofrece la posibilidad de descargar esta tabla en formato CSV, por si resulta de utilidad para el usuario.

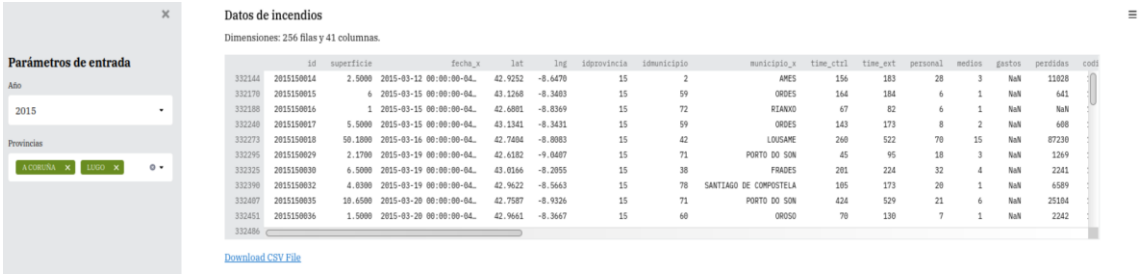


FIGURA 34: PESTAÑA DATOS IV

Por último, se representa la evolución de incendios a lo largo de los meses del año seleccionado, ofreciéndole al usuario la posibilidad de tener una visión real tanto de los focos de incendio como también de los meses de mayor riesgo.

En la Figura 35 se puede observar inicialmente los incendios producidos en el mes de Enero del año 2015 y a continuación se representan los focos ocurridos en el mes de Agosto del mismo año, en ambas ilustraciones observamos la alta concentración en la parte inferior del territorio gallego correspondiente a la provincia de Ourense, además se aprecia que en Agosto se produjo un número mayor de incendios que en Enero, lo que diferencia el riesgo de incendios entre los meses de estudio.

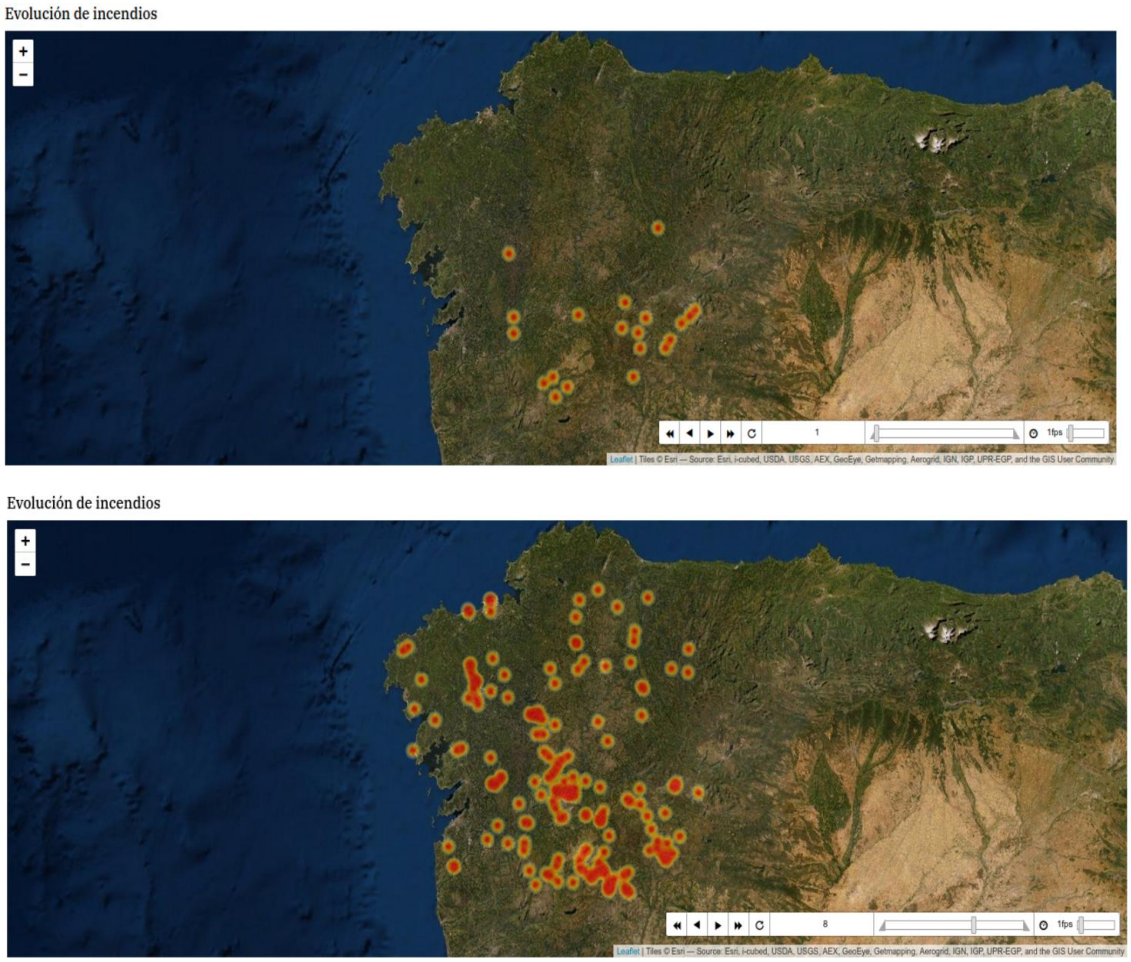


FIGURA 35: PESTAÑA DATOS V

2ª Pestaña: Análisis de Incendios

La segunda pestaña se corresponde con obtención de la probabilidad de incendio utilizando el método explicado en el capítulo 2, aunque se detalló más de un modelo de predicción se utiliza el modelo XGBoost con datos balanceados y con la distinción entre provincias. En esta sección el usuario podrá seleccionar el municipio del que desee conocer la probabilidad de incendio según los datos climatológicos del mes actual (Mayo de 2021), en cuanto seleccione la provincia deseada se le ofrece la lista de municipios correspondientes y tras esta última elección del usuario, nuestro programa se encarga de

localizar la estación meteorológica más próxima a dicho municipio seleccionando los datos climáticos actuales de la misma e introduce dichos datos en el modelo de predicción establecido para dicha provincia, dando como resultado la probabilidad de incendio bajo estas condiciones.

En la Figura 36**¡Error! No se encuentra el origen de la referencia.**, se observa la pestaña en cuestión con una selección por defecto del municipio que se desea analizar. La página además de proporcionar la probabilidad de incendio también muestra la estación más próxima al municipio escogido y los datos climatológicos utilizados para obtener dicha predicción brindando al usuario la posibilidad de una mejor interpretación

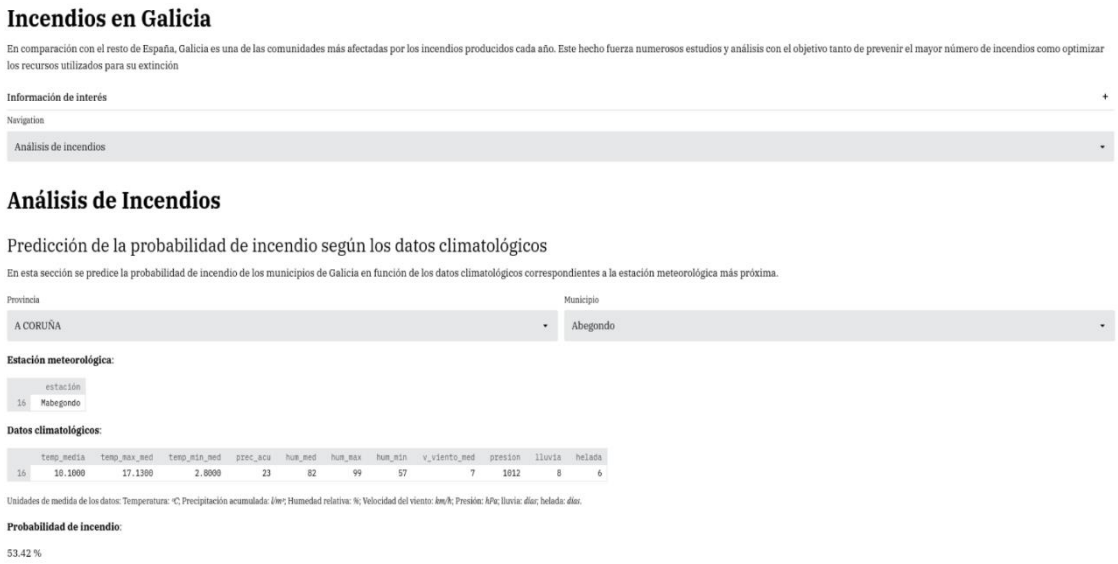


FIGURA 36: PESTAÑA ANÁLISIS DE INCENDIOS

del resultado.

3ª Pestaña: Geolocalización de torres de control

Por último, se introdujo la pestaña correspondiente a la visualización de las localizaciones óptimas de las torres de vigilancias. En ella el usuario podrá seleccionar al número de torres que desea representar (este número lo denotamos por *k*) e internamente nuestro programa utilizará la adaptación del modelo K-means, explicado detalladamente en el capítulo 3, y las localizaciones de los incendios recopilados entre los años 2001 y 2015 para localizar las *k* torres deseadas en puntos óptimos para el control de los incendios en base a este histórico.

En esta sección se representa las localizaciones calculadas para un número determinado de torres conjuntamente a los radios de control de cada una de las torres, adicionalmente a esto, hemos decidido enseñar un mapa con las localizaciones reales de las 44 torres de vigilancia que se encuentran actualmente distribuidas por todo el territorio gallego.

Por defecto, se calcula la ubicación óptima de 44 torres, es decir, se establece k=44, así el usuario puede realizar una comparativa visual entre las localizaciones óptimas y las reales (ver **¡Error! No se encuentra el origen de la referencia.**).

FIGURA 37: PESTAÑA GEOLOCALIZACIÓN DE TORRES DE CONTROL

Conclusión

El objetivo principal de este trabajo era analizar en profundidad ciertos aspectos del problema real que afecta a Galicia año tras año. Aunque hemos realizado este análisis para la predicción de incendios en función de los datos climáticos de cada región y para la ubicación óptima de las torres de vigilancia en función del histórico que poseemos, esto es solo el inicio de un estudio completo, siguiendo con nuestra línea de análisis podríamos utilizar el modelo adaptado de K-Means para la ubicación de parques de bomberos y aeródromos además de realizar un estudio del tiempo de control y extinción en función de la distancia a dichas localizaciones, para poder llegar así a optimizar el tiempo de control de los focos de incendios.

Además, se podría llevar a cabo un análisis de coste y/o perdidas en función de la superficie afectada, la cantidad de medios utilizados, personal, etc. entre otras variables que poseemos (aunque el histórico sea menos para alguna de ellas). Otro análisis interesante, podría ser como de eficiente es la inversión actual comparada con la histórica en prevención de incendios, y como el hecho de mantener los boques limpios y cuidados puede ayudar a la prevención de éstos.

Realmente, creo que con este trabajo hemos logrado enfatizar la seriedad de los incendios sufridos en territorio gallego y que con esfuerzo podríamos llegar a mejorar algunos de los aspectos que se ven afectados por este hecho.

Referencias

Cramer, J. S. (2002). The origins of logistic regression . *Tinbergen Institute*, 167–178.

Embalses. (13 de Mayo de 2021). *Ministerio para la Transición Ecológica, AEMET, SAIH Confederaciones*. Obtenido de https://www.embalses.net/

Garrido, H. (Octubre de 2016). Obtenido de http://hgrosado.carto.com

Hartigan, J. (1975). *Clustering algorithms.* John Wiley & Sons, Inc.

Instituto Galego de Estatística. (s.f.). *Instituto Galego de Estatística*. Obtenido de http://www.ige.eu/web/mostrar_seccion.jsp?idioma=gl&codigo=0101

Instituto Geográfico Nacional. (s.f.). Obtenido de https://www.ign.es/web/ign/portal/ane-datos-geograficos/-/datos-geograficos/datosHidro

Meteogalicia Estaciones. (s.f.). Obtenido de https://www.meteogalicia.gal/observacion/estacions/estacions.action?request_locale=gl#

Meteogalicia Tiempo. (s.f.). Obtenido de https://www.meteogalicia.gal/observacion/estacions/boletins.action

Ministerio de Agricultura, P. y. (s.f.). *CINVIO*. Obtenido de https://datos.civio.es/dataset/todos-los-incendios-forestales/

PLADIGA. (Mayo de 2020). Obtenido de https://mediorural.xunta.gal/sites/default/files/temas/forestal/pladiga/2020/MEMORIA-2020-CAST.pdf