1. **ntroducción**
   1. **Situación actual / descripción del problema**

Hoy en día os incendios forestales suponen un grave problema ecológico, social y económico. Poder saber cuándo y dónde se producen, así como cuál es su extensión, a qué vegetación afectan y, sobre todo, por qué se producen y quién o qué los causa, es algo fundamental para evitarlos, protegernos de ellos y así poder conservar nuestro

Según la organización ecologista internacional ***Greenpeace***, *España es uno de los países de la Unión Europea más afectados por los incendios forestales, un problema que amenaza con intensificarse debido a los efectos del cambio climático.[[1]](#footnote-1)*

En la siguiente imagen se puede ver la cantidad de incendios por Comunidad Autónoma en nuestro país, siendo la zona noroeste la más afectada:

Mapa

Descripción generada automáticamente

A continuación mostramos un gráfico con la evolución de las superficies quemadas en España anuales y acumuladas desde 1961-2016:

**Gráfico, Gráfico de líneas

Descripción generada automáticamente**

*Gráfico 1: Evolución de las superficies afectadas por los incendios forestales en España 1961-2016.*

*Fuente: Observatorio de la Sostenibilidad,* [*https://observatoriosociallacaixa.org/-/incendios-forestales-en-espana-importancia-diagnostico-y-propuestas-para-un-futuro-mas-sostenible*](https://observatoriosociallacaixa.org/-/incendios-forestales-en-espana-importancia-diagnostico-y-propuestas-para-un-futuro-mas-sostenible)

De cada incendio ocurrido en España, por muy pequeño que sea, se debe registrar un Parte de Incendio Forestal. La recopilación de estos datos es de gran utilidad a la hora de elaborar estadísticas para las administraciones forestales, centros de investigación,

Tras analizar todos los datos recogidos de incendios forestales a lo largo de la historia en nuestro país el peor año de la década en cuanto a incendios ha sido 2012, ardiendo un total de 219.956 hectáreas (cuya extensión equivale a la provincia de Vizcaya).[[2]](#footnote-2)

* 1. **Enfoque**

Gracias a la recogida de datos estandarizada se ha podido generar conocimiento sobre el tema, algo totalmente imprescindible a la hora de describir y comprender el fenómeno de los incendios y sus peculiaridades, así como la definición y aplicación de posibles soluciones para afrontarlo.

Es en este punto donde surge la idea de este TFM, queriendo aportar como idea innovadora un modelo que ayude a predecir los incendios, así como las pérdidas económicas que estos puedan suponer.

Crear un modelo que ayude en la predicción de incendios ayudaría a hacer mucho más eficiente la distribución de los recursos necesarios para la extinción y ayudaría también en la reducción de costes, daños y pérdidas. Aunque también hay tener en cuenta la gran dificultad que presenta el desarrollo de un modelo de predicción de incendios, que no es otra que la incertidumbre asociada al comportamiento humano en relación al fuego.

No obstante, se intentará desarrollar un modelo basándonos en el histórico de datos de zonas geográficas con más afluencia de incendios en España.

Una de las Comunidades Autónomas que más sufre este devastador fenómeno año tras año es Galicia. Por ello se ha decidido hacer un enfoque específico al análisis de los incendios producidos en esta Comunidad en este Trabajo Fin de Máster.

*Captura de pantalla de un video juego

Descripción generada automáticamente*A continuación se muestra un análisis de los incendios ocurridos entre los años 2000 y 2015 en Galicia, en el que se puede ver que el número total de incendios producidos en Galicia asciende a un total de 20.543:

*Fuente: elaboración propia*

1. **Sección técnica (análisis profundo de la BBDD)**
   1. **Descripción de los datos**

Para este proyecto se han decidido usar dos BBDD diferentes con el fin de crear una final mucho más completa.

A continuación mostramos un cuadro con todas las variables presentes en nuestro primer dataset, así como una breve explicación de cada una de ellas:

| **Nombre del campo** | **Descripción** |
| --- | --- |
| id | Identificador del incendio |
| superficie | Superficie forestal quemada en hectáreas |
| fecha | Fecha de detección del incendio (formato yyyy-mm-dd) |
| lat | Latitud geográfica del origen del incendio |
| lng | Longitud geográfica del origen del incendio |
| latlng\_explicit | Indica si se dispone de coordenadas geográficas del incendio (1) o se han usado las coordenadas del municipio de origen del incendio (0) |
| comunidad | Identificador de la CC.AA. |
| provincia | Identificador de la provincia |
| municipio | Nombre del municipio |
| causa | Identificador de la causa del incendio |
| causa\_supuesta | Es ‘1’ si la causa es supuesta, si no en blanco |
| causa\_desc | Identificador de la descripción de la causa del incendio |
| muertos | Número de muertos en el incendio |
| heridos | Número de heridos en el incendio |
| time\_ctrl | Tiempo transcurrido hasta entrar en fase de control del incendio (en minutos) |
| time\_ext | Tiempo transcurrido hasta la extinción del incendio (en minutos) |
| personal | Número de personas que han participado en la extinción del incendio (incluye técnicos, agentes forestales, brigadas, bomberos, voluntarios, guardias civiles y ejército) |
| medios | Número de medios terrestres y aéreos que han participado en la extinción del incendio (incluye autobombas, bulldozers, tractores, aviones y otros) |
| gastos | Gastos de extinción asociados al incendio tal y como figuran en EGIF |
| perdidas | Pérdidas económicas asociadas al incendio tal y como figuran en EGIF |

Los valores de algunas variables están incluidos a través de códigos, como son el caso de *comunidad, provincia, causa* y *causa\_dec*. El detalle de estos códigos se encuentra incluido en el apartado Anexos de este proyecto.

Por otro lado, en el segundo dataset se pueden encontrar las siguientes variables:

| **Nombre del campo** | **Descripción** |
| --- | --- |
| Columna 1 | Fecha en formato (aaaa-mm-dd) |
| Columna 2 | Identificador de Estación Meteorológica |
| Columna 3 | Nombre de la Estación |
| Columna 4 | Provincia en la que se encuentra |
| Columna 5 | Altitud sobre nivel del mar (metros) |
| Columna 6 | Temperatura media (ºC) |
| Columna 7 | Precipitación diaria - de 07h a 07 h -(mm = l/m2) |
| Columna 8 | Temperatura minima (ºC) |
| Columna 9 | Hora de Temperatura mínima (hh:mm -UTC-) |
| Columna 10 | Temperatura Máxima (ºC) |
| Columna 11 | Hora de Temperatura máxima (hh:mm -UTC-) |
| Columna 12 | Dirección de la racha máxima (decenas de grado) |
| Columna 13 | Velocidad media (m/s) |
| Columna 14 | Racha (m/s) |
| Columna 15 | Hora Racha (hh:mm -UTC-) |
| Columna 16 | Horas de Sol (horas) |
| Columna 17 | Presión máxima (hPa) |
| Columna 18 | Hora de Presión Máxima (hora entera- redondeada a la hora entera más cercana) |
| Columna 19 | Presión Mínima (hPa) |
| Columna 20 | Hora de Presión Mínima (hora entera- redondeada a la hora entera más cercana) |

* 1. **EDA / calidad de los datos**

La primera BBDD elegida para este TFM ha sido descargada de CIVIO (<https://datos.civio.es/dataset/todos-los-incendios-forestales/>), y sus datos proceden de Estadística General de Incendios Forestales (EGIF).

Por otro lado, la segunda BBDD ha sido descargada de Aemet – Open Data: [**https://opendata.aemet.es/centrodedescargas/productosAEMET**](https://opendata.aemet.es/centrodedescargas/productosAEMET)

* 1. **Programas a usar:**

Para la elaboración del trabajo se ha decidido utilizar 4 programas:

- \*Excel\*: para la fusión de la base de datos que contenía información sobre los incendios forestales en España y la que reunía la información sobre las variables geográficas / geológicas. Se ha utilizado esta opción el lugar de concatenar con otros programas, porque resultaba mucho más visual.

- \*R\* : para el preprocesado y preparación de los datos de los datos. Si bien se han incluido un par de gráficos sobre los valores extremos, se ha optado por utilizar otros programas más óptimos para la representación.

- \*Python (desde Jupyter notebook)\*: dado que se ha profundizado bastante en la programación del máster en la creación de modelos, selección de variables y predicciones a través de esta herramienta, se ha optado por seguir en la misma línea y aplicar así los conocimientos en distintos lenguajes.

-\*ESRI\*: algunos de los componentes del grupo habían trabajado con anterioridad con este programa especializado en la representación de mapas, así que han compartido sus conocimientos con el resto y se ha aprovechado para enriquecer el trabajo añadiendo este programa, con más posibilidad de visualización. Además, el tema elegido para este trabajo requería de una buena visualización que se ha podido plasmar a través de este programa específico.

-\*GitHub\*: se ha usado este portal para compartir el código entre todos los miembros del equipo y así poder hacer cambios.

-\*Streamlit\*: esta herramienta se ha usado para crear una aplicación web en la que se muestra el análisis de los datos usados en este TFM.

Preprocesado Trabajo Fin de de Máster

15/09/2021

## ÍNDICE

### ESTRUCTURA:

1. IMPORTACIÓN DE LIBRERÍAS
2. CREACIÓN DE DIRECTORIO DE TRABAJO E IMPORTACIÓN DE LOS DATOS
3. DESCRIPCIÓN GENERAL DEL DATASET : INSPECCIÓN ANALÍTICA Y VISUAL
4. TRANSFORMACIÓN Y CREACIÓN DE VARIABLES
5. ESTUDIO Y TRATAMIENTO DE OUTLIERS
6. ESTUDIO Y TRATAMIENTO DE VALORES PERDIDOS

### 1. IMPORTACIÓN DE LIBRERÍAS

Vamos a importar las librerías que utilizaremos para la limpieza y el tratamiento de los datos.

### 2. CREACIÓN DE DIRECTORIO DE TRABAJO E IMPORTACIÓN DE LOS DATOS

Tenemos dos alternativas para la lectura de los datos: fijando un directorio de trabajo o leyendo los dataset desde una ubicación web

Lectura de datos desde github ###

### 3. DESCRIPCIÓN GENERAL DEL DATASET : INSPECCIÓN ANALÍTICA Y VISUAL

En este script tenemos contamos con 2 datasets.El de “datos” reúne la información del conjunto del territorio español y el de “datos\_galicia” en el que se va a fundamentar este proyecto. Este dataset es una combinación de dos anteriores, la parte relativa a Galicia del de datos y un conjunto de variables geográficas y geológicas añadidas a posteriori.

#### 3.1) Exploración inicial y tipo de datos / variables y transformación

## [1] 33

Tenemos un total de 33 variables, de las cuales 14 están en el formato de character, 16 en formato de entero y 3 en formato numérico.

En un primer vistazo vemos que el tipo de formato de estas variables lo tendríamos que transformar, para poder trabajar mejor con ellas.

En este sentido vamos a transformar la variable de fecha en el formato date y las variables geográficas y geológicas en numéricas:

También hemos considerado eliminar las variables de id concatenado, latlng\_explicit y causa\_supuesta, porque por la descripción de su contenido la relevancia en nuestro análisis va a ser muy reducida.

Revisamos de nuevo las numéricas para ver si alguna más podría ser transformada en factor,y realizaremos esta transformación en el apartado siguiente

En la transformación de variables que haremos a continuación , cambiará el contenido de la variable y el tipo de datos. En algunos casos la realizaremos para mejorar la interpretabilidad del modelo y para evitar la pérdida de observaciones en el tratamiento de valores extremos y de valores perdidos.

## [1] 25

## [1] 2

### 4. TRANSFORMACIÓN Y CREACIÓN DE NUEVAS VARIABLES

Vamos a hacer una revisión rápida del contenido de cada una de las variables :

Por lo observado, podría ser significativo realizar las siguientes tranformaciones:

**1 . Transformación de idprovincia, municipio y comunidad**

Esta variable vamos a categorizarla porque consideramos que va a ser más informativa de esta forma y por su contenido no tiene sentido tratarla como si fuera numérica.

De igual manera vamos a tratar la variable del municipio.

**2 . Transformación de causas a categórica**

En la variable objetivo categórica simplificamos dos tipologías de causa en una sola ya que la información que reúnen es la misma.

**3 . Creación de nuevas columnas en fecha para : trimestre, años y meses**

Se van a crear 3 varias temporales a partir de “fecha” para mejorar la intrepretabilidad.

**4 . Tranformación de las variables de tiempo**

Se cambiará el formato de las dos variables temporales de minutos a horas sobretodo por cuestiones de interpretabilidad, ya que es más sencillo de valorar en el momento la gravedad del incendio si la duración y el tiempo de extinción se presentan en horas y minutos que en el total de minutos.

**5 . Dicotomizar muertos heridos**

Se van a dicotimizar estas variables asumiendo que en las observaciones donde se dispone de información recogida el valor va a ser 0.

**6 .Categorizar la dirección del viento en función de 8 ejes**

Esta variable se va a tramificar asignándole una orientación en función de los grados que tenga. Hay 8 categorías posibles.

**7 . Dicotomizar precipitaciones**

Esta variable tiene un número de observaciones cuyo valor es cero, muy elevado. Puede ser más significativo el hecho en sí de que las haya o de que no las haya, por eso se va a dicotimizar.

**8 .Sustituir presión máxima y mínima por una columna con su rango**

Se ha creado una nueva variable que sintetiza la información de dos variables de presión máxima y presión mínima en una sola , que es el rango.

**9 . Categorizar los gastos**

A raíz del la función summary inicial, se ha visto que la variable de gastos tenía una proporción importante de valores perdidos y parecía razonable categorizarla.

### 5. ESTUDIO Y TRATAMIENTO DE OUTLIERS - VALORES EXTREMOS

#### 5.1 Identificación de los outliers

Podemos utilizar algunas funciones de apoyo que hemos creado:

Se va a crear un subset de variables numéricas para representar un primer boxplot con las transformaciones que tenemos.Para ello, la columna de id se fija como nombre de las filas y la variable objetivo de la causa se elimina.De este dataset se extraen las numéricas que se representarán.

Dentro de input vamos a hacer un subset para las numéricas y otro para las categóricas porque la identificación es distinta:

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

En la mayoría de variables es evidente que hay valores extremos y a raíz del análisis gráfico, recibirán tratamientos distintos:

Para cada una de las variables contenidas en el subset de numéricas input, vamos a identificar los valores extremos así como calcular su incidencia. Después crearemos el dataframe de incidencia a partir de este, que contendrá en una columna los nombres de las variables y en otra su incidencia.

En este dataset se realizará un tramiento parcial de los valores atípicos ya que algunos casos pueden llegar a ser considerados puntos de interés . Por ejemplo, en el caso de la superficie quemada,la presencia de valores anómalos es fuerte en esta variable, ya que un porcentaje significativo de las observaciones se corresponde con extensiones muy elevadas de terrenos quemado. La presencia de incendios con estas grandes dimensiones, aunque sea reducida en su conjunto es significativa, por eso no podemos omitirla en su totalidad.

#### 5.2 Tratamiento de los outliers

En primer lugar se ha visto que en la variable de altitud el porcentaje de extremos es muy alto por lo que se va a categorizar para evitar tener que eliminar un 12,27 % de las observaciones. En este caso la transformación a variables categóricas va a ser una tramificación

**10 . Tramificar altitud**

Después de realizar esta transformación tenemos que volver a cargar la parte correspondiente de las variables numéricas para conocer la incidencia de las variables restantes como numéricas.

Se comienza con la eliminación de los valores extremos de las variables de medios, superficie, el tiempo transcurrido hasta entrar en fase de control del incendio ( time\_ctrl) y el tiempo transcurrido hasta la extinción del incendio ( time\_ext)

Una vez eliminados estos outliers, se realizará otra transformación de la variable de VELMEDIA, ya que el porcentaje de valores extremos (9.16%) sobre el total de las observaciones es significativo, por lo que se va a categorizar.

**11 Categorizar velocidad media**

Se continúa con la eliminación de los valores extremos, en este caso para las variables de latitud y del rango entre la presión máxima y la presión mínima y la variable de personal.

Se regenera el listado de incidencia en el que se puede observar que la presencia de valores extremos se ha reducido significativamente y solo queda una indicendia residual en algunas de las variables que se mantiene por si fueran puntos de interés.

Se representa el boxplot para ver cómo ha quedado finalmente la distribución de los valores extremos:

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

No se van a tratar los outliers en la variable de pérdidas porque a partir de la funcións de summary inicial se ha visto el elvado número elevado de observaciones perdidas y muy posiblemente sea eliminada esta variable.

### 6. ESTUDIO Y TRATAMIENTO DE VALORES PERDIDOS

En esta sección se vela por que la eliminación de los valores sea efectiva, ya que no proporcionan información e impiden la correcta aplicación de los modelos al dataset.

Ahora se identificarán los valores faltantes:

#### 6.1 Localización

Vamos a comprobar la cantidad de valores perdidos por cada una de las variables en valor absoluto y porcentual

En las variables de pérdidas, de velocidad media y de dirección del viento hay valores faltantes. Solo son 3 variables a tratar porque el tratamiento de valores extremos ya ha eliminado muchas filas.

Las posibilidades del tratamiento son las mismas que para los outliers: la eliminación de las filas o la imputación de estos valores por la media o la mediana.

* En el caso de la variable de pérdidas se eliminarán en vez de imputar ya que el rango es demasiado grande para imputar.

\*En el caso de las variables categóricas de dirección del viento y de velocidad media, de igual manera es muy arriesgado situar todas las observaciones perdidas en una de las categorías, porque distorsionaría el dataset.

Vamos a eliminar las observaciones en estas 3 variables que están afectadas y comprobaremos de nuevo la presencia de valores perdidos.

Eliminamos la variable de pérdidas por el alto % de NAs

## [1] 13422 28

El dataset tratado cuenta con X líneas

write.csv(datos\_galicia, file = "datos\_galicia\_limpio.csv")

**MODELADO**

Después de la preparación de datos, tenemos que desarrollar el modelado.

Un modelo tiene como objetivo predecir una variable objetivo a partir de un conjunto de variables explicativas y lo vamos a desarrollar mediante técnicas de Machine Learning siguiendo los siguientes pasos:

1. EDA
2. División de nuestro dataset
3. Modelos
4. Mejora de modelos
5. Selección del modelo

Para esta parte del trabajo, vamos a utilizar lenguaje Python a través de Júpiter Notebook.

1. **EDA - Análisis exploratorio de datos**

Partimos de nuestra base de datos limpia de R. Un dataset con un total de 12976 líneas y 29 columnas

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Las columnas son variables que debemos saber qué tipo de datos tienen y cómo se distribuyen; esto lo haremos viendo el tipo de datos, su histograma y el número de observaciones únicas de cada variable.

Tabla

Descripción generada automáticamente con confianza media

Gráfico de cajas y bigotes

Descripción generada automáticamente

Tabla

Descripción generada automáticamente con confianza mediaInterfaz de usuario gráfica, Tabla

Descripción generada automáticamente con confianza media

A simple vista, lo que vemos, es que las variables Unnamed e id tienen tantos valores únicos como observaciones contienen, por lo que decidimos eliminar ambas variables.

Por otro lado, las variables se dividen entre variables numéricas y variables categóricas. En Python no podemos trabajar con variables categóricas, por lo que tendremos que hacer un tratamiento especial con ellas y por ello, vamos a dividir el dataset en dos:

Texto

Descripción generada automáticamente con confianza bajaTexto

Descripción generada automáticamente

En primer lugar, eliminaremos aquellas que tengan demasiadas categorías, como Idmunicipio y fecha, que tienen un total de 268 y 2272 valores únicos respectivamente. Al tener tantas categorías y estar representada de una manera menos fraccionada en idprovincia para el caso de idmunicipio y Mes, Trimestre y Año para fecha, decidimos prescindir de ellas.

Las formas más comunes de tratar este tipo de variables suelen ser transformarlas a variables ordinales o realizar la codificación One-Hot.

Transformarlas en orden numérico en ocasiones dificulta la predicción ya que da diferentes pesos a las distintas categorías de una variable, por ello, esta transformación la vamos a utilizar para aquellas variables que sí sigan un orden en sus categorías; sin embargo, para aquellas que no sigan un orden, utilizaremos la codificación One-Hot. Este método consiste en crear una nueva variable binaria por cada categoría existente en la variable inicial, donde 1 serán las observaciones que pertenezcan a esa categoría y 0 las demás.

Para saber qué variable será transformada con cada método, vemos qué categorías son las que componen nuestras variables.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamenteTexto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamenteImagen que contiene Texto

Descripción generada automáticamente

Tabla

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamente con confianza baja

* **variables con codificación One-Hot**: idprovincia, Trimestre, Mes, DIR\_VIENTO
* **variables orden numérico**: gastos, ALTITUD, VELMEDIA

Tras realizar estas transformaciones y previa comprobación de que no se han traspasado los datos con ningún valor faltante, nuestro dataset tiene un total de 12976 líneas y 45 columnas y está listo para comenzar con el modelo.

1. **DIVISIÓN DE NUESTRO DATASET**

En primer lugar, antes de empezar a generar nuestros modelos, debemos definir nuestra variable objetivo, que será causa. El resto de variables serán nuestras variables explicativas.

Recordemos cómo se distribuye nuestra variable, ya que lo deberemos tener en cuenta a la hora de seleccionar el tipo de modelo que queremos realizar.

Escala de tiempo

Descripción generada automáticamente con confianza media

La variable causa se compone de cinco categorías, las cuales están desbalanceadas y debere- mos tenerlo en cuenta.

En segundo lugar, tenemos que dividir de manera aleatoria nuestro dataset en dos, una parte que usaremos para entrenar nuestro modelo que corresponderá con el 80% a la que llamaremos Train y otra para evaluarlo, con el 20% restante que será Test. En esta parte es en la primera en la que deberemos tener en cuenta el desbalanceo de nuestra variable objetivo, distribuyendo de manera proporcional las categorías.

Texto

Descripción generada automáticamente

Una vez realizados estos dos pasos, ya podemos comenzar a generar nuestros modelos.

1. **MODELOS**

Para los modelos vamos a utilizar métodos Backward (Eliminación hacia atrás), partimos de modelos iniciales muy complejos que incluyen todas las variables y después, vamos reduciéndolas.

Al querer predecir una variable multicategoría, utilizaremos los modelos que consideramos pueden ser más idóneos para ella.

* Support Vector Machines = SVM : Consiste en un conjunto de métodos de aprendizaje supervisados que realiza un método one-against-one . Utiliza una técnica de iguala los dintintos tipos de categorías, por lo tanto, es una buena opción para predecir una variable categórica desbalanceada.
* Árbol de decisión: Combinación y subdivisión en ramas de las variables de una forma binaria para la toma de la decisión con mayor probabilidad de que ocurra un suceso
* Random Forest: Conjunto de árboles de decisión
* Gradient Boosting: Está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. La diferencia con Random Forest es que utiliza árboles más débiles, con menos profundidad.

Estos modelos no serán normalizados ya que hemos comprobado que no afecta al resultado final.

Las métricas de rendimiento que vamos a utilizar es el Acurracy, utilizando una escala del 0 al

1, en la que 0 indica una capacidad nula de predicción y 1 que el modelo tendría la capacidad

de predecir correctamente el 100% de las causas de los incendios.

Los resultados de los modelos más complejos con todas nuestras variables iniciales, son los siguientes:

'MODELO 1 :Score modelo SVM : 0.8701848998459168 con todas las variables: 44'

'MODELO 2 : Score modelo Árbol de decisión : 0.7619414483821263 con todas las variables: 44'

'MODELO 3 : Score modelo Random Forest : 0.8701848998459168 con todas las variables: 44'

'MODELO 4 : Score modelo Gradient Boosting : 0.8701848998459168 con todas las variables : 44 '

Para empezar, tenemos unos resultados muy buenos, parece que los modelos se ajustan muy

bien; ahora, nuestro siguiente paso, es reducir las variables para poder conseguir un modelo más sencillo y un mejor rendimiento.

Al ser el Árbol de decisión el modelo que peores resultados nos da en una primera valoración

vamos a prescindir de él.

1. **MEJORA DE MODELOS**

Para mejorar nuestro modelo vamos a realizar los siguientes ajustes:

* Selección de Parámetros
* Selección de variables

**4.1 SELECCIÓN DE PARÁMETROS**

Para ello, hemos utilizado Gridsearch,una técnica que selecciona los parámetros que debemos utilizar para sacar un mayor rendimiento a nuestro modelo indicándole el modelo que queremos utilizar y los parámetros a probar.

**4.2 SELECCIÓN DE VARIABLES**

Iremos reduciendo variables a nuestro dataset a través de tres métodos de filtro:

4.2.1.Teniendo en cuenta la correlación entre variables predictoras y la correlación de esas variables con la variable objetivo

Mostramos la correlación entre las variables que superen el 50% de correlación entre ellas y las variables que tengan una correlación superior al 5% con la variable objetivo.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamenteGráfico

Descripción generada automáticamente

Eliminamos en primer lugar las variables más correlacionadas entre si (TMAX,TMIN y time\_ext) una vez eliminadas, nos quedamos con las variables que estén por encima del 5% de correlación con la variable objetivo.

Los modelos con esta selección de variables dan los siguientes resultados:

'MODELO 5 : Score modelo SVM reducido teniendo en cuenta correlaciones : 0.8701848998459168 con 13 variables'

'MODELO 6 :Score modelo Random Forest reducido teniendo en cuenta correlaciones : 0.8713405238828967 con 13 variables'

'MODELO 7 : Score modelo Gradient Boosting reducido teniendo en cuenta correlaciones : 0.8701848998459168 con 13 variables'

4.2.2. Mejores variables por el método de chi-cuadrado: Evalúa la probabilidad de correlación entre las variables.

Este método selecciona las siguientes variables : ['superficie', 'lat','lng','muertos','heridos','time\_ctrl','time\_ext','personal','medios', 'PRECIPITACION']

Los resultados de los modelos serían los siguientes:

'MODELO 8 : Score modelo SMV por Chi-cuadrado : 0.8701848998459168 con 10 variables'

'MODELO 9: Score modelo SMV por Chi-cuadrado : 0.8713405238828967 con 10 variables'

'MODELO 10 : Score modelo Gradient Boosting por Chi-cuadrado : 0.8701848998459168 con 10 variables'

### 4.2.3. Selección de las mejores variables a través la importancia de las variables predictoras

Este método solo se ha realizado con Random Forest y Gradient Boosting ,ya que el modelo SVM no lo permite.

Las variables seleccionadas y los resultados para cada uno de ellos serían los siguientes:

**['lat', 'lng', 'personal', 'medios', 'TMAX', 'Año']**

'MODELO 11 : Score modelo Gradient Boosting respecto a las variables con mayor importancia : 0.8701848998459168 con 7 variables'

**['superficie', 'lat', 'lng', 'time\_ctrl', 'personal', 'medios', 'TMEDIA','RACHA', 'SOL', 'Año', 'PRES\_RANGE']**

'MODELO 12 : Score modelo Random Forest respecto a las variables con mayor importancia : 0.8713405238828967 con 12 variables'

1. **SELECCIÓN DEL MODELO**

Cuando realizamos las predicciones en Train, podemos contrastarlas con los datos reales en Test. Si los valores son cercanos, quiere decir que nuestro modelo lo podemos dar por bueno y para evaluarlo utilizaremos la validación cruzada.

La validación cruzada selecciona distintos cortes (en este caso hemos seleccionado 5) para realizar el entrenamiento y la validación del modelo comprobando que los resultados obtenidos son iguales independientes a la muestra.

Tras realizar la validación cruzada a los modelos que consideramos más óptimos, vamos a representar este método mediante boxplot para los tres modelos que hemos ido utilizando.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Los tres modelos tienen una capacidad predictiva por encima del 85% y varianzas muy bajas, lo que demuestra que sus predicciones no son aleatorias. El grafico junto a los resultados de la validación cruzada de los modelos concretos nos lleva a seleccionar el modelo que consideramos que llega a obtener los mejores resultados, así como su sencillez. Por tanto, el modelo seleccionado es Random Forest con la selección de variables del MODELO 12, cuyos resultados de la validación cruzada son los siguientes:

**Texto

Descripción generada automáticamente**

Además, podemos ver también la matriz de confusión de nuestro modelo, así como las variables que lo componen y su importancia.

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

En la matriz de confusión vemos que en la primera clase de la variable predice prácticamente el 100% de los valores, sin embargo, en el resto de categorías sucede todo lo contrario, lo cual seguramente sea causado por el excesivo desbalanceo con el que contábamos. Vamos a comprobarlo mediante un reporte de clasificaciones donde nos indica los resultados de las predicciones de manera desglosada:

Tabla

Descripción generada automáticamente

Efectivamente, las predicciones de la 3º,4º y 5º clase son muy bajas.

Tanto la matriz de confusión como los resultados de nuestras clases indican que el desbalanceo afecta a los resultados, ya que el Accuracy se consigue debido a que la mayoría de la precisión viene dada por nuestra clase mayoritaria y por lo tanto, en este caso , quizás no sea la mejor manera de medir nuestro modelo; para poder realizar una predicción más precisa realizamos un método por el cual se generan observaciones adicionales en las clasificaciones minoritarias, con ello, nuestros resultados serían:

Tabla

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

Se reduce la precisión de la primera categoría, pero mejora la precisión del resto.

\*Futuras mejoras: realizar un modelo dicotómico, separando la variable únicamente entre incendios intencionados y no intencionados. Visión de negocio: aseguradoras.

1. Fuente: *Greenpeace International*, <https://es.greenpeace.org/es/trabajamos-en/bosques/incendios-forestales/> [↑](#footnote-ref-1)
2. Fuente: *Greenpeace International*, <https://es.greenpeace.org/es/trabajamos-en/bosques/incendios-forestales/> [↑](#footnote-ref-2)