Índice general

1. Análisis exploratorio de datos **3**

1.1. Distribución de la variable objetivo 4

1.2. Análisis univariantes de las variables numéricas 6

1.3. Análisis multivariantes de las variables numéricas 11

1.4. Análisis de las variables categóricas 13

Bibliografía **15**

Capítulo 1

Análisis exploratorio de datos

En este capítulo se aplicarán distintos métodos numéricos y gráficos de análisis de datos a la muestra generada siguiendo el procedimiento detallado en el capítulo anterior. Se usarán principalmente técnicas de estadística descriptiva para comprender las características del conjunto de datos y extraer información útil para el problema que se intenta abordar, predecir incendios forestales. Es importante tener presente que se trata de datos correlados espacial y temporalmente, lo que hace necesario el uso de métodos específicos para este tipo de datos. Los objetivos de esta etapa son:

1. Generar conocimiento sobre el conjunto de datos que nos permita evaluar la calidad de este, sin olvidar las limitaciones que ya se han comentado en la sección anterior.
2. Conocer, al menos de forma descriptiva, el impacto de cada variable en la variable objetivo. Este conocimiento será necesario para evaluar e interpretar los modelos que se construirán en la próxima sección.
3. Analizar las características de las distintas variables, de cara a usar posteriormente técnicas de preprocesamiento adecuadas para cada modelo.

Antes de abordar el estudio detallado de cada una de las variables y las relaciones entre estas, en la Figura 1.1 se recoge un resumen de todo el conjunto de datos, sin incluir la columna de geometría. En este resumen se puede observar que en el conjunto de datos hay 4 tipos de variables (además de la variable *geometry* que es de tipo *simple feature column POINT*, abreviado como *sfc_POINT*): cadenas de caracteres, fechas, factores y variables numéricas.

Se puede observar que hay registros en 749 municipios diferentes (de los 785 municipios de que hay en Andalucía). Probablemente el hecho de que en algunos municipios no haya habido observaciones sea debido a los datos faltantes. Las variables *municipio* y *cod_municipio* no se incorporarán a los modelos. De la misma forma, se puede ver que hay observaciones en 3691 días diferentes.

El conjunto cuenta con 5 variables de tipo factor: *fire* (la variable objetivo), *WD10M*, *orientacion*, *enp* y *uso_suelo*; y con 18 variables numéricas. Aunque cada una de ellas se analizará a continuación con detalle, ya cabe hacer algunos comentarios:

- El 38 % de las observaciones se encuentran en espacios de vegetación arbustiva y/o herbácea (código 32).

- Como era de esperar, por la forma en la que se ha tomado la muestra, el conjunto está balanceado.
- El 81 % de las observaciones se encuentran fuera de Espacios Naturales Protegidos.
- Todas las variables, salvo *T2M* y *curvatura*, son positivas y la mayoría de ellas presentan una marcada distribución asimétrica hacia la derecha.
- Las variables muestran escalas muy diversas entre ellas, siendo *GWETTOP* la que presenta menor desviación típica (0.145) y *poblacion* la que tiene una desviación típica mayor (64453). Se evidencia la necesidad de incluir algún método de normalización de las variables en el preprocesamiento de los datos.

```
── Data Summary ────────────────────────
                           Values
Name                       datos
Number of rows             21546
Number of columns          26
_____________________
Column type frequency:
  character                2
  Date                     1
  factor                   5
  numeric                  18
_____________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 cod_municipio         0             1   5   5     0      749          0
2 municipio             0             1   3  32     0      749          0

── Variable type: Date ───────────────────────────────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min        max        median     n_unique
1 date                  0             1 2002-01-02 2022-11-29 2012-08-04     3691

── Variable type: factor ─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 fire                  0             1 FALSE          2 1: 10794, 0: 10752
2 WD10M                 0             1 FALSE          8 SW: 4965, W: 4867, S: 3786, SE: 3316
3 orientacion           0             1 FALSE          9 S: 3267, SW: 3090, SE: 2956, W: 2544
4 enp                   0             1 FALSE          2 0: 17393, 1: 4153
5 uso_suelo             0             1 FALSE         15 32: 8068, 21: 3128, 24: 2798, 22: 2786

── Variable type: numeric ────────────────────────────────────────────────────────────────────────────────────────────────────────────────────
   skim_variable    n_missing complete_rate      mean        sd      p0      p25      p50      p75     p100 hist
 1 T2M                      0             1     24.6       5.20    -0.8     22.7     25.6     28.0     36.0 ▁▁▃▇▂
 2 GWETTOP                  0             1      0.318     0.145    0.09     0.2      0.29     0.42     0.91 ▇▇▆▂▁
 3 RH2M                     0             1     46.4      16.1      9.23    33.8     44.6     57.5     96.8 ▂▆▇▆▂
 4 WS10M                    0             1      3.64      1.42     0.99     2.67     3.34     4.25    13.3  ▇▇▂▁▁
 5 PRECTOTCORR              0             1      0.229     1.38     0        0        0        0       46.1  ▇▁▁▁▁
 6 elevacion                0             1    494.      400.       0      169.     425.     711.    3351.  ▇▃▂▁▁
 7 pendiente                0             1     12.3      11.8      0        3.46     8.24    17.9     98.0  ▇▃▂▁▁
 8 curvatura                0             1      0.00549   0.0563  -0.294   -0.0193  -0.000100  0.0255  0.420 ▁▁▇▂▁
 9 dist_carretera           0             1   1799.     1907.       0.0706  507.     1200.    2425.    17662. ▇▃▂▁▁
10 dist_poblacion           0             1    902.      800.       0        379.     713.     1170.     8215. ▇▃▂▁▁
11 dist_electr              0             1   5253.     5640.       0.220   1216.    3462.    7276.    48069. ▇▂▁▁▁
12 dist_ferrocarril         0             1  15311.    13521.       0.824   5086.   11855.   21467.    85242. ▇▅▂▁▁
13 dist_camino              0             1    762.      741.       0.0238   240.     539.     1060.     7090. ▇▃▂▁▁
14 dist_sendero             0             1   5619.     4805.       0.0465 1781.     4361.    8276.    25103. ▇▆▃▂▁
15 poblacion                0             1  22095.    64453.     114      2252     4860     16759    704414 ▇▁▁▁▁
16 dens_poblacion           0             1    105.      320.       2.30    12.2     30.8     71.1     4974. ▇▁▁▁▁
17 dist_rios                0             1   6781.     5836.       1.94   2162.    5237.    10123.    37074. ▇▅▂▁▁
18 NDVI                     0             1      0.412     0.136    0        0.314    0.393    0.495    0.944 ▁▃▇▃▁
```

Figura 1.1: Resumen numérico del conjunto de datos depurados. *Fuente: Elaboración propia empleando la función* `skim` *del paquete "skimr" [2].*

1.1. Distribución de la variable objetivo

En primer lugar, se estudiará la distribución de la variable *fire* espacial y temporalmente.

En la Figura 1.2 se muestran los histogramas de la variable objetivo en función del día de la semana, del mes y del año, respectivamente. En primero el de ellos se observa que, mientras que la distribución de los casos negativos es uniforme entre los días de la semana, en los casos positivos se aprecia un ligero aumento en el fin de semana, especialmente en el

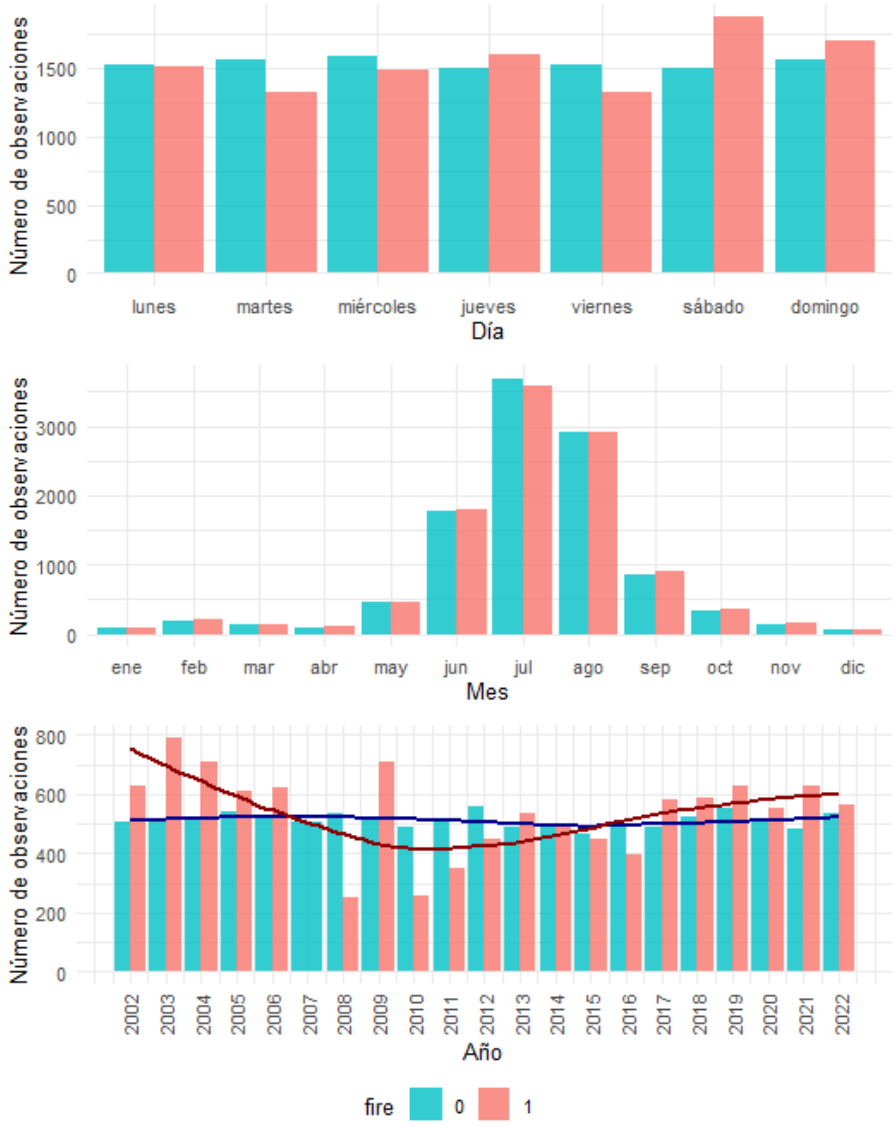

Figura 1.2: Distribución temporal de la variable objetivo. *Fuente: Elaboración propia.*

sábado. Esto podría ser algo meramente casual o deberse al hecho de que más personas van al campo durante el fin de semana por motivos de ocio y se producen más desplazamientos, lo que podría podría aumentar el riesgo de que se produzcan negligencias o accidentes que desencadenen incendios forestales (en 2022, el 39.23 % de las actuaciones forestales registradas fueron debidas a neglicencias u accidentes [1]). En el segundo histograma, se observa como las observaciones se concentran en los meses de verano y en cada mes hay una cantidad balanceada de muestras de ambas clases (esto es fruto del proceso de muestreo de las observaciones negativas, que como se ha explicado en la sección anterior, se ha llevado acabo asegurando que la proporción de casos negativos en cada mes sea igual a la de los casos positivos). En el tercer histograma es remarcable que, mientra las observaciones negativas están uniformemente distribuidas entre los 20 años del estudio, las positivas muestran una disminución importante en los años 2008 y 2010. En el año 2007 no hay observaciones positivas, debido a que los 4 polígonos de incendios mayores de 100 *ha* que había registrados ese año no disponían de la fecha de inicio del incendio, por lo que no pudieron usarse para el estudio. Se desconoce la causa del reducido número de incendios (mayores de 100 *ha*) en 2007, 2008 y 2010.

Dada la clara influencia del mes y la aparente influencia del día de la semana en la aparición de incendios (al menos, con certeza, de aquellos de la dimensión correspondiente al estudio), estas variables serán incluidas en los modelos a través del procesamiento de la variable *date*. Al hacer esto se está hipotetizando que el efecto del mes o del día de la semana va más allá de las características meteorológicas propias de cada periodo, ya que también contiene información sobre otras dimensiones, como las diferentes tendencias sociales durante el periodo vacacional o los cambios en los movimientos de población durante el fin de semana o las distintas estaciones. Dada la imposibilidad de medir todas estas variables individualmente, se espera que al menos parte del efecto que puedan tener sobre la aparición de incendios forestales quede recogido a través de su estacionalidad.

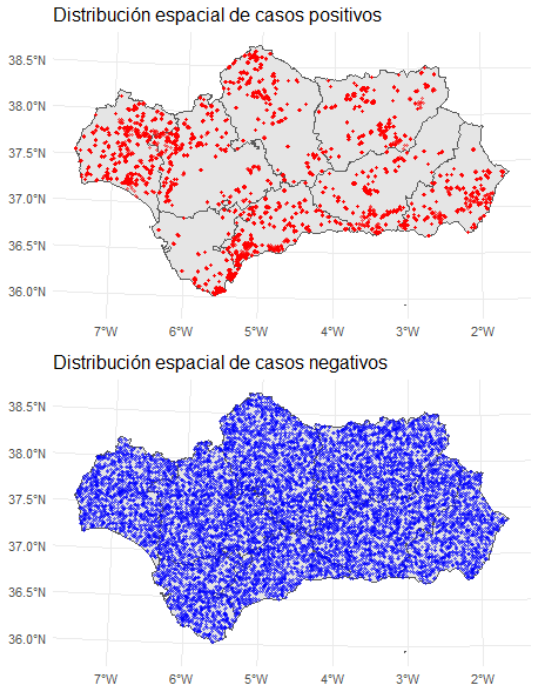

Figura 1.3: Distribución espacial de la variable objetivo. *Fuente: Elaboración propia.*

En la Figura 1.3 se observa claramente cómo las 10.752 muestras negativas están uniformemente distribuidas dentro de los límites de la Comunidad Autónoma de Andalucía, mientras que las 10794 muestras positivas se concentran a ambos lados de la cuenca del río Guadalquivir, con una mayor densidad de observaciones en la provincia de Huelva y en algunas zonas de la costa mediterránea (como ya se apreciaba en la Figura **??**).

1.2. Análisis univariantes de las variables numéricas

El análisis univariante de las variables numéricas se lleva a cabo desde 3 enfoques complementarios:

1. A través de la los resúmenes numéricos recogidos en la Figura 1.1 y del análisis gráfico de los diagramas de caja y bigotes (Figura 1.4).
2. Estudiando la media mensual de cada variable en función de la variable *fire*.

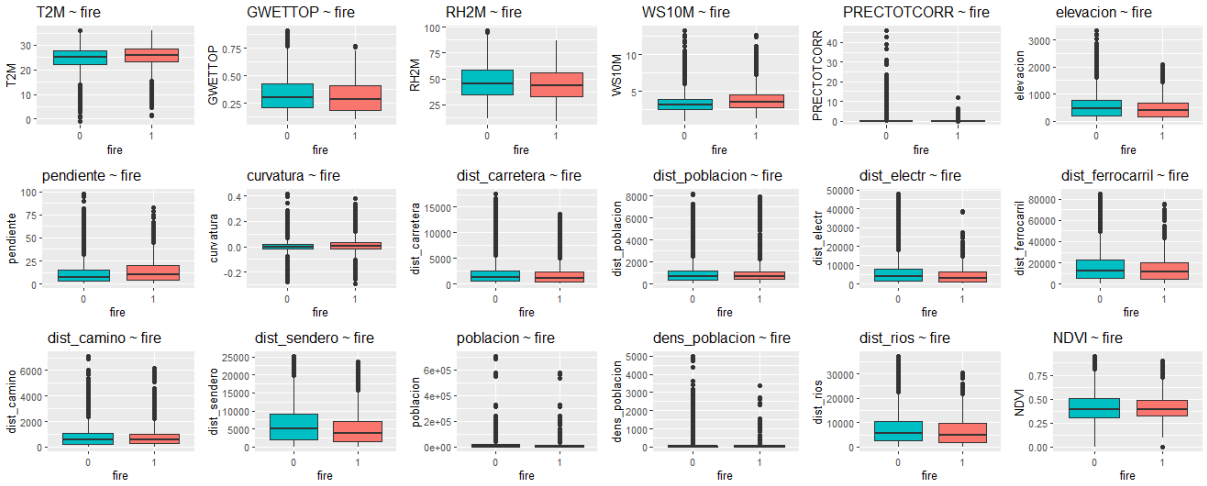

Figura 1.4: Boxplot de cada variable numérica en función de la variable objetivo. *Fuente: Elaboración propia.*

3. Analizando la distribución espacial de cada variable separando por mes si corresponde. Los gráficos correspondientes a este análisis se recogen en el [Apéndice A.1][Gráficos espaciales EDA].

En los *boxplots* de las variables numéricas en función de la variable *fire* (Figura 1.4) destacan varios aspectos. Por un lado, como ya se había comentado anteriormente, las variables presentan escalas muy diferentes y la mayoría tiene una marcada asimetría hacia la derecha. Por otro lado, es evidente la gran cantidad de valores *outliers* que se observan en los datos, lo que tendrá implicaciones en los modelos que se construyan con ellos. Sin embargo, es importante destacar que no se trata de observaciones erróneas, sino que son inherentes a la naturaleza de los datos. Por ejemplo, en el caso de la variable *PRECTOTCORR* el valor máximo observado es 46.06 *mm* en un día, un valor elevado que sin duda es atípico en esta región de clima seco, pero sin embargo, posible. Es también remarcable que todas las variables presentan una variabilidad similar en ambos niveles del factor *fire*, lo que indica que no será un problema de clasificación trivial. *A priori*, solo con los diagramas de caja y bigotes y los resúmenes numéricos es difícil llegar a más conclusiones, sin embargo, sí pueden observarse sutiles diferencias entre las distribuciones de algunas variables para ambos niveles del factor *fire*.

Dada la naturaleza temporal de algunas variables, el análisis gráfico de los *boxplots* resulta insufiente. Con el fin de considerar la componente estacional de las variables climáticas y de vegetación, se estudiará, a continuación, la media mensual de cada una de estas variables en función de la variable objetivo.

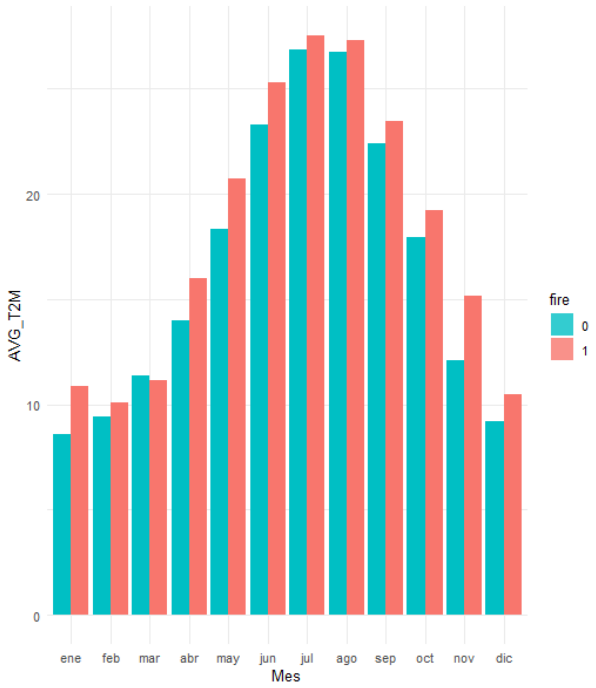

Figura 1.5: Media mensual de $T2M$ en función de $fire$. *Fuente: Elaboración propia.*

En la Figura 1.5 se puede observar cómo en casi todos los meses, la temperatura media mensual es superior en las observaciones en las que se ha registrado un incendio forestal.

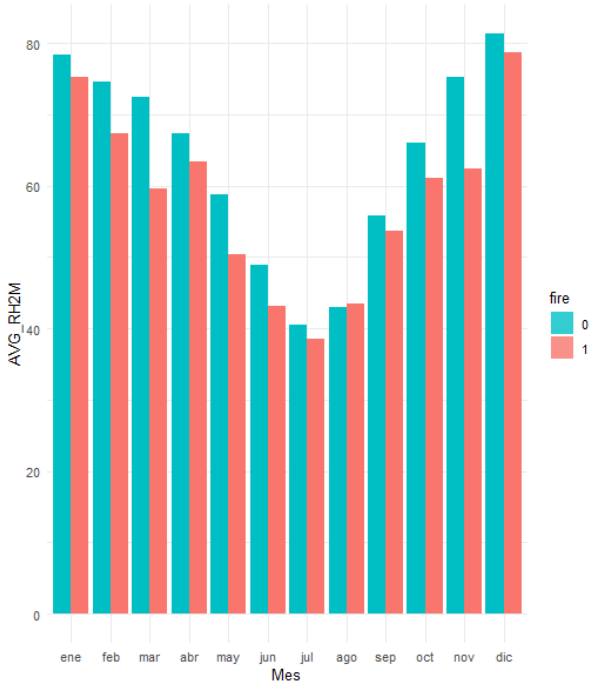

Figura 1.6: Media mensual de $RH2M$ en función de $fire$. *Fuente: Elaboración propia.*

En la Figura 1.6 se puede observar que en todos los meses la media mensual de la humedad relativa del aire a 2 m sobre la superficie es menor en las observaciones en las que se ha registrado un incendio forestal. Sin embargo, las diferencias se reducen durante los meses de verano, en los que la humedad presenta valores bajos en ambas clases.

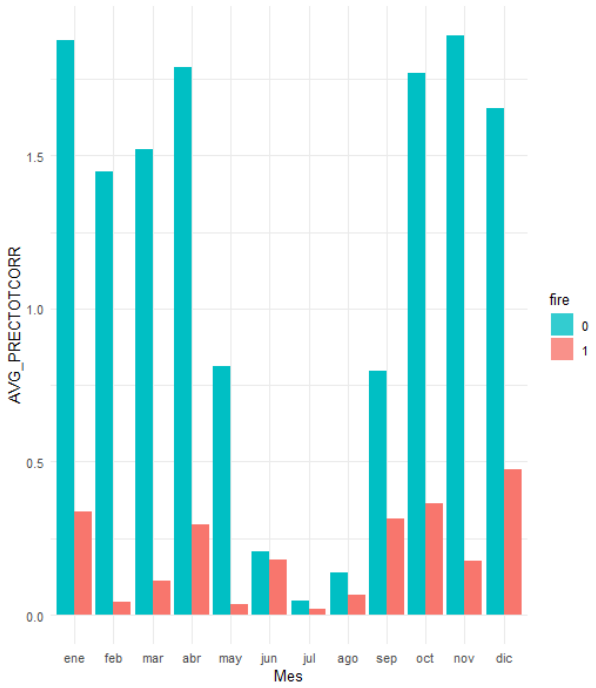

Figura 1.7: Media mensual de $PRECTOTCORR$ en función de $fire$. *Fuente: Elaboración propia.*

En la Figura 1.7 se observa una clara diferencia en la media mensual de las precipitaciones diarias en función de si se ha registrado o no un incendio forestal en la observación, siendo significativamente mayor en este último caso.

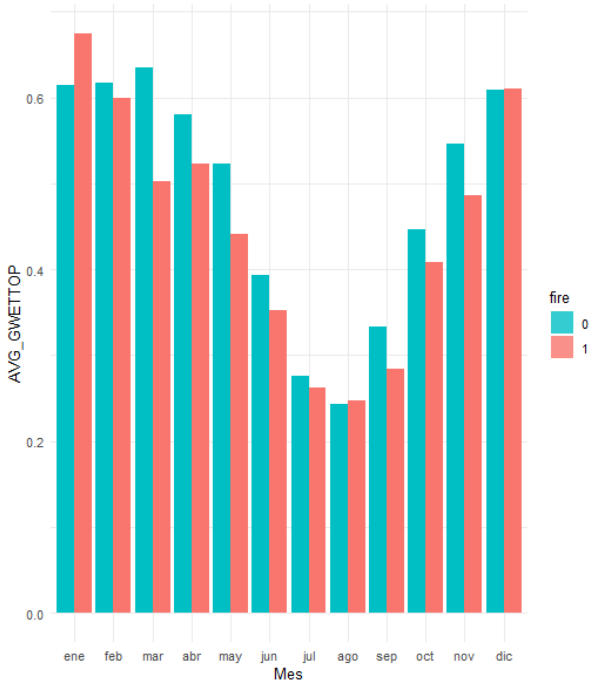

Figura 1.8: Media mensual de $GWETTOP$ en función de $fire$. *Fuente: Elaboración propia.*

En la Figura 1.8, que muestra la media mensual de la humedad del suelo en función de

si en esa observación se ha registrado o no incendio, se observa un gráfico similar al de la humedad relativa del aire, con valores medios más elevados en las observaciones en las que no se han registrado incendio forestal. Sin embargo, también parece que las diferencias son más reducidas durante la estación estival.

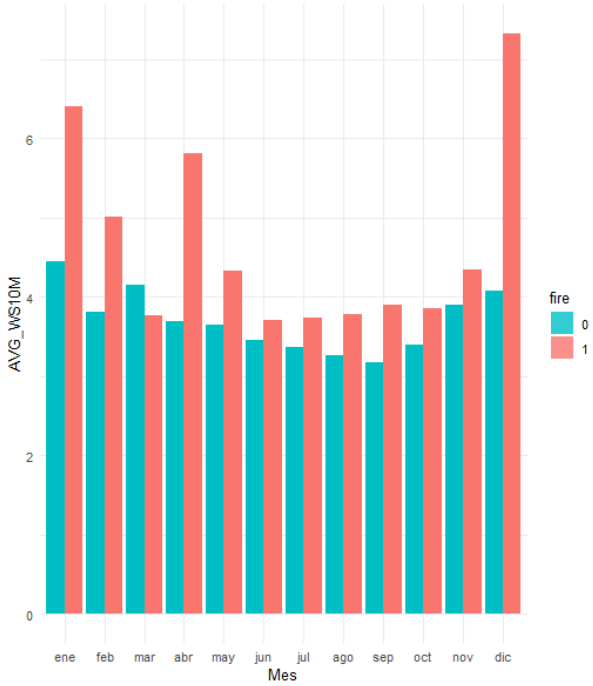

Figura 1.9: Media mensual de $WS10M$ en función de $fire$. *Fuente: Elaboración propia.*

En la Figura 1.9 se observa cómo, durante todos los meses, la media mensual de la velocidad del viento a 10 metros sobre la superficie es mayor en los registros en los que ha habido un incendio forestal. Esto es, de hecho, algo remarcable. Ya que los incendios que se están considerando son incendios que han llegado a calcinar un área superior a 100 ha, es de esperar que se trate de incendios que se dan bajo unas condiciones meteorológicas concretas que facilitan la rápida propagación del fuego, dificultado su extinción temprana. Y dado que se está considerando la fecha de inicio de los incendios, esto es lo que podría estar viéndose en la Figura 1.9.

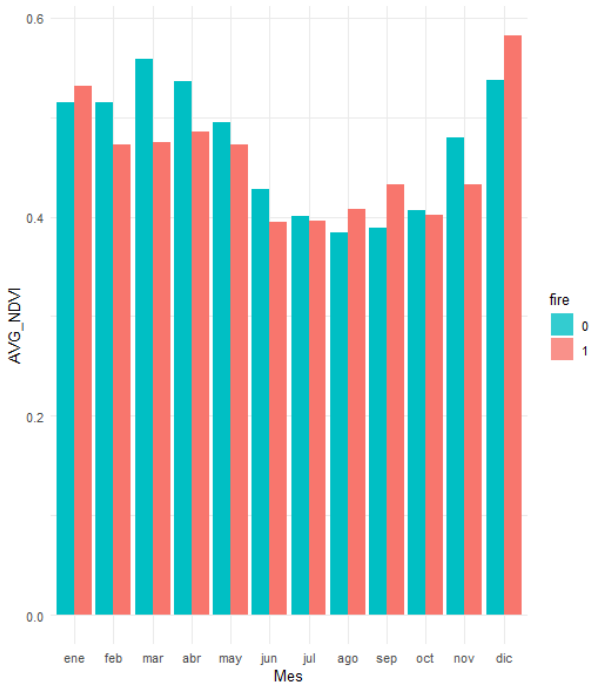

Figura 1.10: Media mensual de $NDVI$ en función de $fire$. *Fuente: Elaboración propia.*

Como se observa en la Figura 1.10, las diferencias entre los casos en los que se ha registrado incendio y los que no, en términos del *NDVI*, no están claras. Se recuerda que esta variable se utiliza para cuantificar la cantidad y verdor de la vegetación, por lo que es coherente que se observen valores medios más bajos en los meses de verano.

En el [Apéndice A.1][Gráficos espaciales EDA] se recogen los gráficos espaciales y espacio-temporales de todas las variables numéricas. En ellos se refleja cómo los valores de las variables en estudio son coherentes con lo que cabría esperar de la realidad. Además, permiten una comprensión mayor de la distribución espacial (y temporal) de las variables en el área de estudio, lo que será útil de cara a interpretar los modelos que se construyan.

1.3. Análisis multivariantes de las variables numéricas

En la Figura 1.11 se muestra un gráfico con las correlaciones entre las variables. La interpretación es sencilla, cuanto más intenso sea el color y cuanto mayor sea la excentricidad de la elipse, mayor será la correlación (en valor absoluto) para ese par de variables. El color de la elipse indica el signo del coeficiente de correlación. De esta forma, se observa que las variables más correlacionadas en la muestra son: *T2M* con *RH2M* (negativamente, -0.71), *T2M* con *GWETTOP* (negativamente, -0.69), *GWETTOP* con *R2HM* (positivamente, 0.68) y *poblacion* con *dens_poblacion* (positivamente, 0.63). Esto es razonable, ya que cabe esperar que al aumentar la temperatura del aire disminuya la humedad del aire y del suelo; que al aumentar la humedad del suelo aumente también la del aire y viceversa; y que municipios muy densamente poblados también tengan un elevado número de habitantes. Cabe destacar que no se trata de valores alarmantes como para considerar *a*

priori que las variables proporcionan información redundante, susceptible de ser eliminadas en un paso de ingeniería de características. Sin embargo, sí se probará a reducir la dimensionalidad de los datos capturando la mayor parte de la varianza a través de PCA.

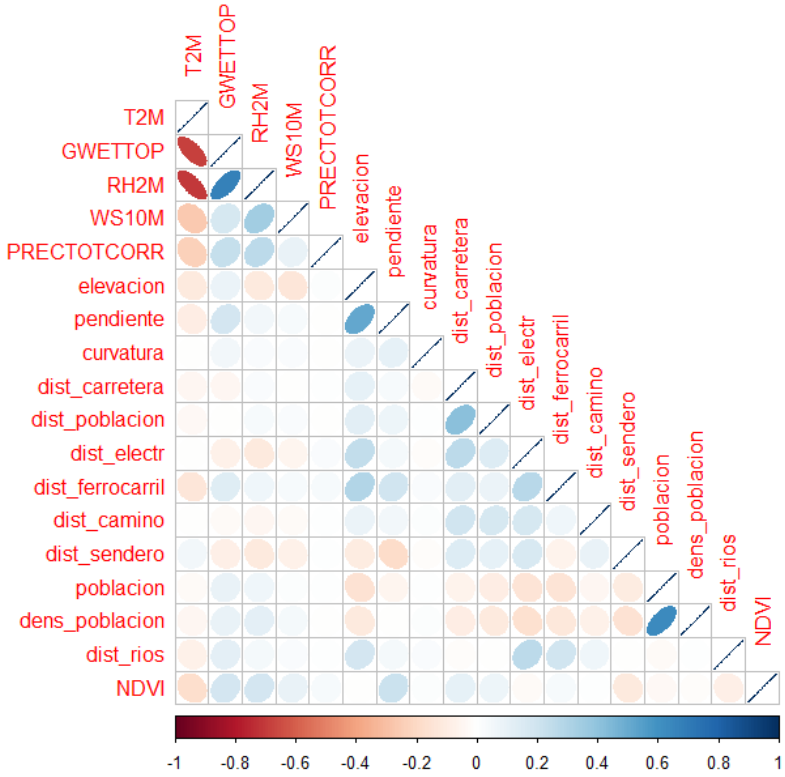

Figura 1.11: Correlaciones entre variables numéricas. *Fuente: Elaboración propia.*

En la Figura 1.12 se muestra el gráfico de coordenadas paralelas de las variables tipificadas a una normal estándar, es decir, restándoles la media y dividiendo por la desviación típica. Este gráfico complementa la información de los *boxplots*, pues refleja también las relaciones entre las variables. Si bien es cierto que al tener un número bastante elevado de observaciones el gráfico no es tan claro, pueden hacerse algunas observaciones importantes.

En primer lugar, se observa que la variable con mayor variabilidad (una vez tipificada) es PRECTOTCORR, que presenta bastantes valores atípicos, todos ellos en observaciones en las que no se ha registrado incendio. También destacan en este sentido *dens_poblacion* y *poblacion*, entre las que además puede observarse que no hay una relación lineal clara (hay municipios con un elevado número de habitantes pero con una densidad de población reducida y viceversa). Además, puede verse que todas las variables tienen una marcada asimetría positiva (salvo *curvatura*, *T2M* y *NDVI*). Este gráfico es útil también pues permite ver a qué clase de *fire* corresponden los valores más atípicos de cada variable. Por ejemplo: la mayor parte de los valores más elevados de *WS10M*, *dist_poblacion*, *curvatura* y *dist_camino* se dan en observaciones positivas, mientras que en *PRECTOTCORR*, *elevacion*, *GWTTOP*, *dist_Carretera*, *dist_electr* y *dist_rios* sucede lo contrario.

Los resultados de aplicar análisis de componentes principales sobre la matriz de correlaciones de las 18 variables numéricas se muestran en la Figura 1.13. Como se puede observar, se necesitan al menos 11 componentes principales para lograr explicar el 80 % de la varianza de la muestra, y 14 para alcanzar el 90 % de la varianza de los datos. Estos

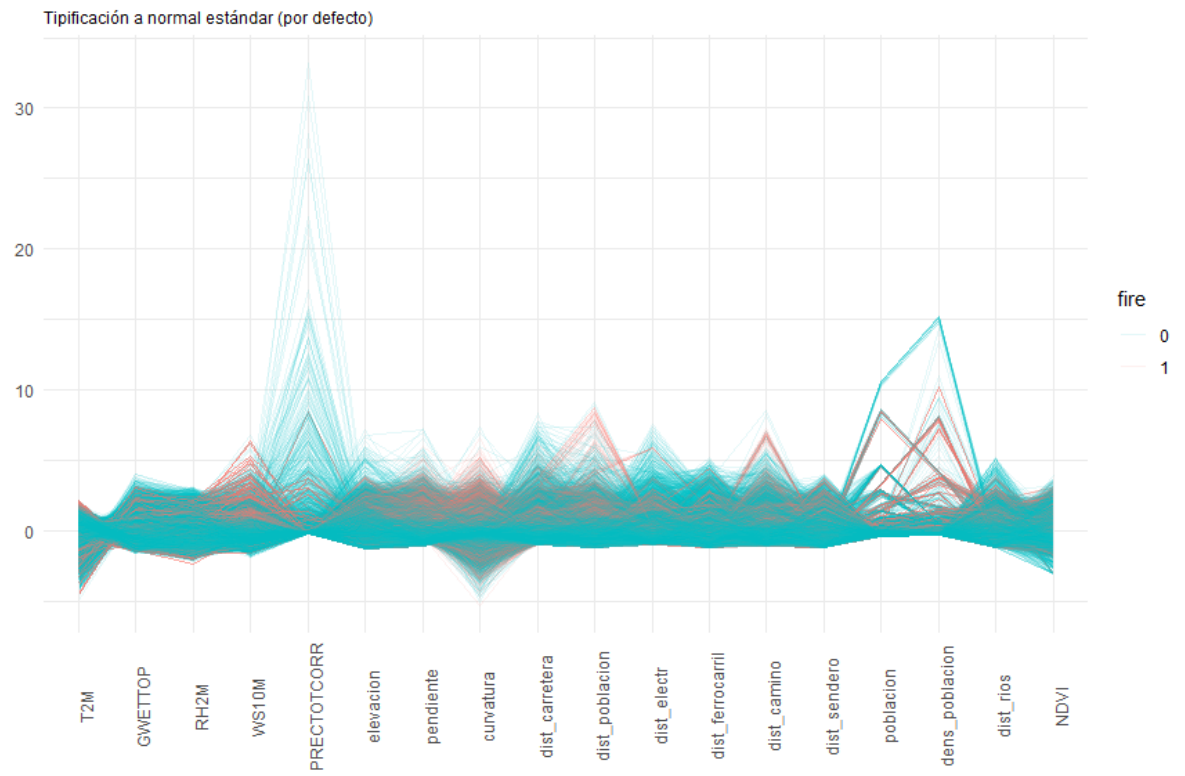

Figura 1.12: Gráfico de coordenadas paralelas de las variables numéricas tipificadas. *Fuente: Elaboración propia.*

resultados se aplicarán más adelante en los modelos, pero a nivel meramente explicativo ya indican que se trata de un conjunto de datos complejo en cuanto a la dimensión real de estos.

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8    Comp.9
Standard deviation     1.6830901 1.5509109 1.26704790 1.18114361 1.13860921 1.00151777 0.98498153 0.92948032 0.92853341
Proportion of Variance 0.1573773 0.1336291 0.08918946 0.07750557 0.07202394 0.05572433 0.05389937 0.04799631 0.04789857
Cumulative Proportion  0.1573773 0.2910065 0.38019595 0.45770152 0.52972546 0.58544979 0.63934916 0.68734547 0.73524404
                          Comp.10    Comp.11    Comp.12    Comp.13    Comp.14    Comp.15    Comp.16    Comp.17    Comp.18
Standard deviation     0.91036573 0.86805852 0.85739620 0.78965012 0.72570700 0.65041713 0.60872397 0.52343451 0.48001944
Proportion of Variance 0.04604254 0.04186253 0.04084046 0.03464152 0.02925837 0.02350236 0.02058583 0.01522132 0.01280104
Cumulative Proportion  0.78128658 0.82314912 0.86398958 0.89863109 0.92788946 0.95139182 0.97197765 0.98719896 1.00000000
```

Figura 1.13: PCA sobre la matriz de correlaciones de las variables numéricas. *Fuente: Elaboración propia.*

1.4. Análisis de las variables categóricas

Las variables categóricas se analizarán a través de los histogramas de cada variable en función de la variable *fire* (Figura 1.14).

En la variable *WD10M* cabe destacar la escasez de observaciones con dirección del viento norte. En el histograma no se observa una clara relación de esta variable con la variable objetivo, aunque entre las observaciones con viento con dirección sur o suroeste hay más observaciones negativas y entre las que tienen dirección noroeste o este hay una mayor presencia de observaciones positivas.

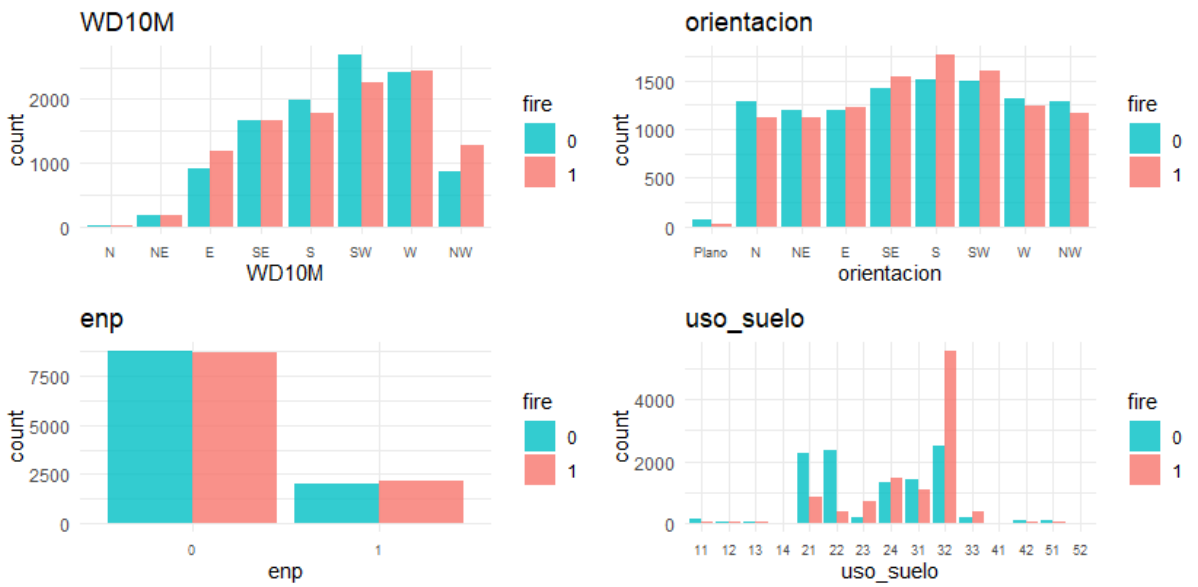

Figura 1.14: Histogramas de las variables categóricas en función de $fire$. *Fuente: Elaboración propia.*

En el caso de la variable *orientación*, la relación tampoco está clara, aunque puede verse una mayor proporción de observaciones positivas en las superficies con orientación sur (sureste, sur y suroeste).

En términos de la variable *enp* por si sola no se observan diferencias significativas entre ambas clases.

La variable *uso_suelo* sí que muestra una distribución marcadamente diferenciada entre ambas clases. La mayoría de las observaciones positivas se dan en espacios de vegetación arbustiva y/o herbácea (código 32), clase en la que hay casi el doble de observaciones positivas que negativas. En tierras de labor (código 21) y cultivos permanentes (código 22) la proporción de observaciones negativas es mucho mayor, mientras que en zonas agrícolas heterogéneas (código 24) y en espacios abiertos con poca o sin vegetación (código 33) hay una mayor presencia de observaciones positivas dentro de la muestra. También es relevante el hecho de que casi la totalidad de las observaciones se encuentran en zonas agrícolas y en zonas forestales (que se corresponden con los códigos comenzados por 2 o 3, respectivamente), mientras que en las demás clases la proporción de observaciones es mucho menor (3.5 % de total). Es por ello que antes de construir los modelos, todas las categorías de uso de suelo que no se corresponden con zonas agrícolas o forestales (es decir, todas cuyo código no comienza por 2 o 3) se agruparán en el nivel *Otro*. En la Figura **??** del [Apéndice A1][Gráficos espaciales EDA] puede observarse la distribución espacial de esta variable.

Prueba [Gráficos espaciales EDA][]

Bibliografía

[1] (2023). «Datos Estadísticos Andalucía del 01/01/2022 al 31/12/2022 Plan INFOCA, Centro Operativo Regional». *Informe técnico*, Consejería de Sostenibilidad, Medioambiente y Economía Azul, Junta de Andalucía.
https://www.juntadeandalucia.es/medioambiente/portal/documents/20151/55229821/memoria-infoca-Andalucia-2022.pdf/b95604c3-53d7-7564-bbb3-6c8a4f8c332a?t=1713950214344.

[2] Waring, Elin; Quinn, Michael; McNamara, Amelia; Arino de la Rubia, Eduardo; Zhu, Hao y Ellis, Shannon (2022). *skimr: Compact and Flexible Summaries of Data*.
https://docs.ropensci.org/skimr/(website). R package version 2.1.5, https://github.com/ropensci/skimr/.