

APC - Informe Práctica 1

Lidia Conde Matos(1570710), June Gorostidi Elizetxea(1566312),
Jordi Calbet Escandell(1564567)

1 Apartado C

Los incendios forestales pueden crear problemas ecológicos y poner en peligro vidas humanas y propiedades. Entender cuándo ocurren y cuáles son sus causas es importante para controlarlos. Los datos con los que trabajaremos en este proyecto están asociados con un papel de investigación científica sobre la predicción de casos de incendios forestales en Portugal usando técnicas de modelización.

Estas son las variables del dataset y el rango de valores para cada uno:

X: axis X del mapa del parque Montesinho: 1 a 9

Y: axis Y del mapa del parque Montesinho: 2 a 9

month: mes del año: 'jan' a 'dec'

day: día de la semana: 'mon' a 'sun'

FFMC: índice de 'Fine Fuel Moisture Code' del sistema Fire Weather Index (FWI): 18.7 a 96.2

DMC: índice de 'Duff Moisture Code' del sistema FWI: 1.1 a 291.3

DC: índice de 'Drought Code' del sistema FWI: 7.9 a 860.6

ISI: índice de 'Initial Spread' del sistema FWI: 0.0 a 56.1

temp: temperatura en grados Celsius: 2.2 a 33.3

RH: humedad relativa en porcentaje: 15 a 100

wind: velocidad del viento en km/h: 0.4 a 9.4

rain: lluvia en mm/m2: 0.0 a 6.4

area: área quemada del bosque en hectáreas: 0 a 1090.84

1.1 Información del dataset

La siguiente tabla nos da información sobre nuestro dataset: de qué tipo es cada atributo, el número de entradas, memoria usada... Como se puede apreciar, en

nuestro dataset no hay ningún valor nulo. Gracias a esto no necesitaremos modificarlo para trabajar con él.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 517 entries, 0 to 516
Data columns (total 13 columns):
 #   Column  Non-Null Count  Dtype  
--- 
 0   X        517 non-null    int64  
 1   Y        517 non-null    int64  
 2   month    517 non-null    int64  
 3   day      517 non-null    int64  
 4   FFMC    517 non-null    float64 
 5   DMC     517 non-null    float64 
 6   DC      517 non-null    float64 
 7   ISI     517 non-null    float64 
 8   temp    517 non-null    float64 
 9   RH      517 non-null    int64  
 10  wind    517 non-null    float64 
 11  rain    517 non-null    float64 
 12  area    517 non-null    float64 
dtypes: float64(8), int64(5)
memory usage: 52.6 KB
```

Por otro lado, en esta tabla aparecen las estadísticas de los atributos numéricos del dataset. Entre ellos la media, la desviación estándar y los cuartiles. Estos datos se podrán observar mejor con los boxplots de más adelante, relacionados con los meses del año.

Mostramos las estadísticas de los atributos numéricos de la BBDD:

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
count	517.000	517.000	517.000	517.000	517.000	517.000	517.000	517.000	517.000	517.000	517.000	517.000	517.000
mean	4.669	4.300	7.476	4.259	90.645	110.872	547.940	9.022	18.889	44.288	4.018	0.022	12.847
std	2.314	1.230	2.276	2.073	5.520	64.044	248.068	4.559	5.807	16.317	1.792	0.296	63.656
min	1.000	2.000	1.000	1.000	18.700	1.100	7.900	0.000	2.200	15.000	0.400	0.000	0.000
25%	3.000	4.000	7.000	2.000	90.200	68.600	437.700	6.500	15.500	33.000	2.700	0.000	0.000
50%	4.000	4.000	8.000	5.000	91.600	108.300	664.200	8.400	19.300	42.000	4.000	0.000	0.520
75%	7.000	5.000	9.000	6.000	92.900	142.400	713.900	10.800	22.800	53.000	4.900	0.000	6.570
max	9.000	9.000	12.000	7.000	96.200	291.300	860.600	56.100	33.300	100.000	9.400	6.400	1090.840

1.2 Ralación entre el código FWI y las condiciones meteorológicas

Después de calcular la media de cada atributo respecto a los meses hemos podido sacar las siguientes observaciones sobre el tiempo:

- Las precipitaciones ocurren en los meses de agosto, julio y marzo (este último en menor cantidad).
- Junio, julio, agosto, septiembre y octubre tienen temperaturas altas.
- El viento es bajo en enero, febrero, julio, septiembre y octubre.
- Vemos que la humedad también es baja en septiembre, octubre, noviembre y diciembre.
- Los valores de DC son altos en julio, agosto, septiembre y más secos en octubre.
- Los valores de DMC son altos en julio, agosto y septiembre, pero no en octubre.
- Los valores de FFMC superan los 90 en agosto, julio, septiembre y octubre.
- Los valores de ISI son altos en julio, agosto y septiembre.
- Los meses que no tengan precipitaciones tienen mayor posibilidad de tener incendios.
- Cuando la temperatura aumenta, el contenido de humedad de los tres tipos

de combustible se reduce en el mismo mes, así que en los meses de julio, agosto, septiembre y octubre aparecen condiciones más peligrosas vistas desde el punto de vista del tiempo.

- También podemos ver que los valores de DMC y DC no tienen relaciones significativas con las columnas del viento y 'Relative humidity.'
- Los meses que tengan unos valores de humedad bajos son más propensos a tener incendios.

1.3 Histograma de los atributos



En el histograma de los meses, es interesante ver que los valores anormalmente altos de fuegos forestales ocurren en los meses de agosto y septiembre. En el caso de los días, los días de viernes a lunes son los que tienen mayor proporción de casos.

A su vez, cuando mas altos son los valores de 'FFMC' y 'DC', más incendios forestales habrá, es decir, tienen una correlación positiva. En cambio, el caso de 'DMC' es mucho mas irregular y tiene mas probabilidad de incendio en los

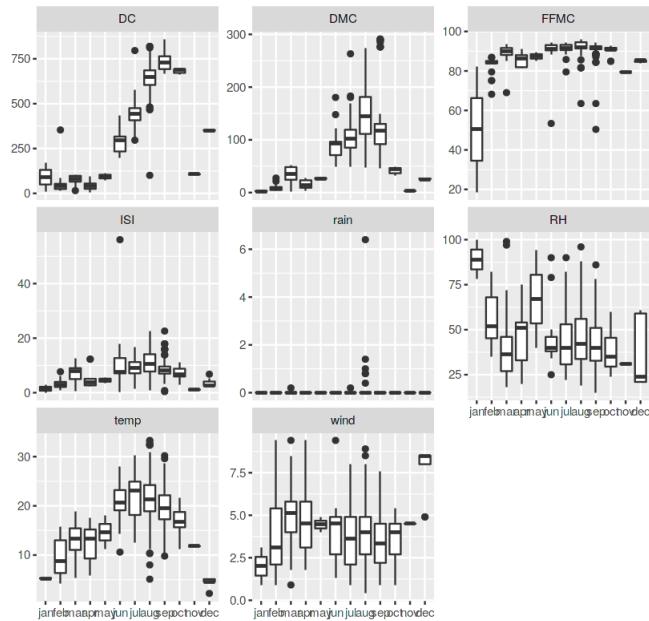
valores entre 100-150.

En cuanto a la temperatura, se puede observar que sigue más o menos una distribución gaussiana y que la mayoría de incendios ocurren a los 15-20 grados y no a partir de los 30, que es lo que cabría esperar. Esto podría ser porque rondando los 20 grados es cuando la gente más sale a la montaña y a causa de esta actividad se producen más incendios.

Por otro lado, se puede apreciar tanto en la 'ISI' como en el humedad relativa que el pico mas alto esta desplazado a la izquierda, es decir, que en valores pequeños, mayores son los incendios, ya que estos valores están opuestamente correlacionados. Asimismo, en el histograma de la lluvia se ve claramente que solo cuando no llueve ocurren los incendios forestales. En cambio, en el histograma del viento se puede ver que hay más incendios más o menos en la mitad, ya que hay suficiente viento para propagarlo pero no lo suficiente para apagarlo.

Para acabar, en el histograma del área se puede observar que los mayores valores del área se encuentran entre el cero y el cien, por lo que hay unos cuantos 'outliers' con valores muy altos que tendremos que tratar para hacer la regresión.

1.4 Gráficos de la relación entre los atributos y los meses



- DC:

Se puede apreciar claramente que los meses de agosto, septiembre y octubre son los que tienen más DC, seguidos de Junio, julio y diciembre que se acercan un

poco. También se puede observar que noviembre y diciembre tienen pocos datos, ya que no varían apenas. Agosto tiene bastantes outliers, pero la mayoría no son demasiado extremos. Para acabar, se puede ver que septiembre y octubre no están distribuidos normalmente.

- DMC:

En el caso del 'DMC' junio, julio, agosto y septiembre son los que más tienen. De los mencionados anteriormente, agosto tiene una variación más alta de los datos y septiembre no está distribuido normalmente. Todos ellos tienen bastantes outliers, sobre todo septiembre, en el que están muy dispersos. En cambio, en el resto de meses no hay casi DMC.

- FFMC:

En este caso, la 'FFMC' es muy alta en casi todos los meses y en consecuencia, los datos no nos dicen mucho, ya que todas son muy parecidos. No obstante, en enero hay una varianza muy grande.

- ISI:

Este caso se parece un poco al anterior pero con los valores bajos. Se puede apreciar que los meses que tienen unos valores más altos son marzo y desde junio a octubre. En este caso, como en el primero, podemos apreciar que septiembre tiene muchos 'outliers', pero tiene menos variación que agosto, por ejemplo. Para acabar, se puede ver que julio tiene una mediana bastante baja.

-rain:

En la gráfica se puede observar que no llueve casi nada, excepto en algunos momentos específicos; sobre todo en agosto y en menor medida en marzo y julio.

-RH:

En esta gráfica se puede observar que enero tiene un nivel anormalmente alto de RH y en cambio marzo y los meses de mayo a diciembre son los que menos tienen.

-temp:

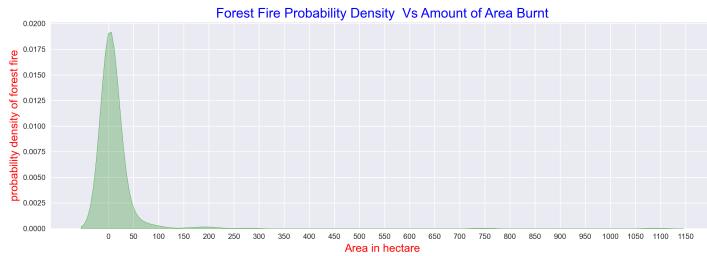
Las temperaturas más altas son en los meses de junio a octubre. Esta vez, agosto tiene muchos outliers y septiembre unos pocos. En este caso todos los meses están normalmente distribuidos.

-wind:

Para empezar, se puede observar que diciembre está muy alto y en cambio, enero está bastante bajo. El resto tienen mas o menos los mismos valores. Los únicos que tienen poca varianza son enero, mayo, noviembre y diciembre.

En conclusión, se puede ver en todos los gráficos que la mayoría de datos están muy elevados en los meses de junio a octubre y se podría llegar a la conclusión de que hay muchos incendios en esos meses.

1.5 Añadiendo variables categóricas según el área quemada



Para empezar, con las primeras gráficas podemos ver que el área quemada está muy sesgada con un valor de +12,84 hectáreas y un enorme valor de curtosis de 194 hectáreas. También nos dice que la mayoría de los incendios forestales no cubren un área muy grande, y que la mayor parte del área dañada está por debajo de 50 hectáreas de terreno.

Por otra parte, en las otras gráficas podemos observar que un número anormalmente alto de incendios forestales ocurren en el mes de agosto y septiembre y al menos en noviembre, y que en el caso de día, los días viernes, sábados, domingos y lunes tienen mayor proporción de casos. No hay un indicador sólido, pero pensamos que podría ser debido a que los fines de semana la gente sale más de su casa, por lo cual, a más gente más posibilidad de que haya un incendio.

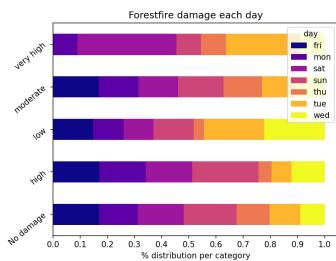


Figure 1: Gráfico 1.

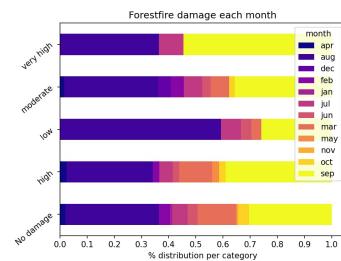
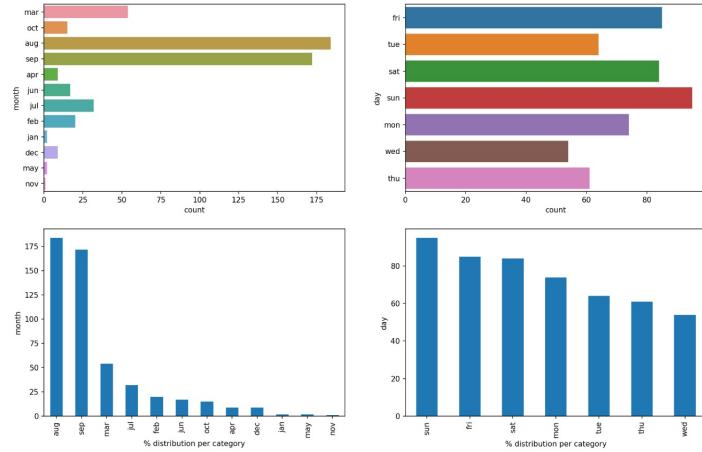
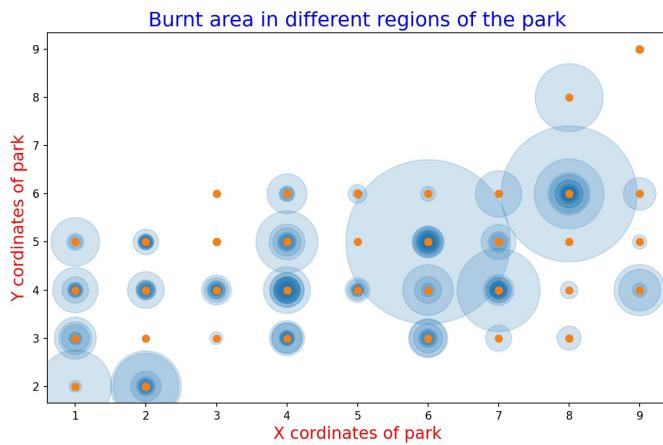


Figure 2: Gráfico 2.

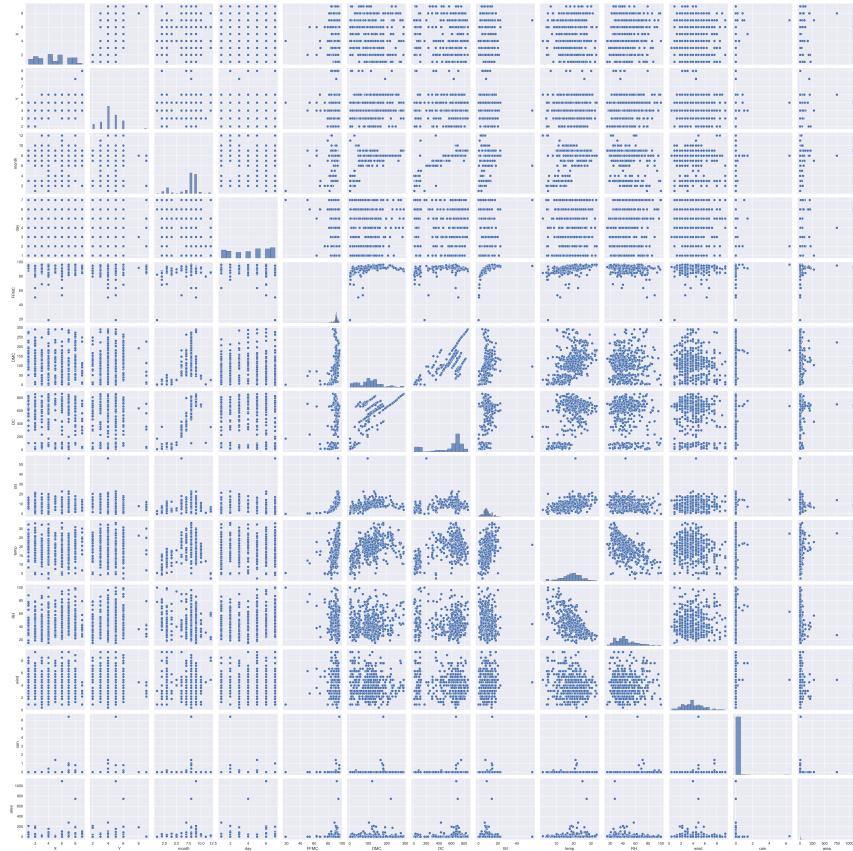
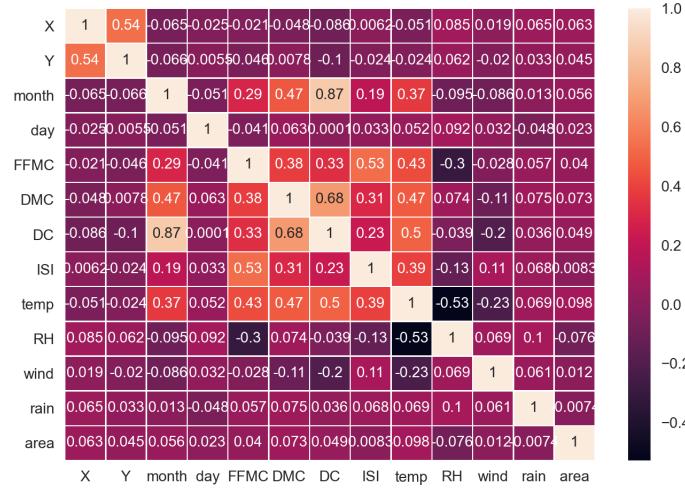


1.6 Análisis del área quemada según las coordenadas espaciales (X,Y)



En esta gráfica, podemos ver como representamos el parque en un eje de 9x9, y representamos las zonas donde se han producido incendios por sus pesos. Podemos observar que la coordenada (6,5) muestra la zona quemada más intensa, seguida de la coordenada (8,6).

1.7 Correlación



En la tabla, podemos ver como cada celda nos dice el valor de la correlación entre las dos variables implicadas. Este varía en el intervalo [-1,1], estableciendo el signo el sentido de la relación, y la interpretación de cada resultado es el siguiente:

- Si $r = 1$: Correlación positiva perfecta. El índice refleja la dependencia total entre ambas variables, la que se denomina relación directa: cuando una de las variables aumenta, la otra variable aumenta en proporción constante.
- Si $0 < r < 1$: Refleja que se da una correlación positiva.
- Si $r = 0$: En este caso no hay una relación lineal. Aunque no significa que las variables sean independientes, ya que puede haber relaciones no lineales entre ambas variables.
- Si $-1 < r < 0$: Indica que existe una correlación negativa.
- Si $r = -1$: Indica una correlación negativa perfecta y una dependencia total entre ambas variables lo que se conoce como "relación inversa": una de las variables aumenta, la otra variable en cambio disminuye en proporción constante.

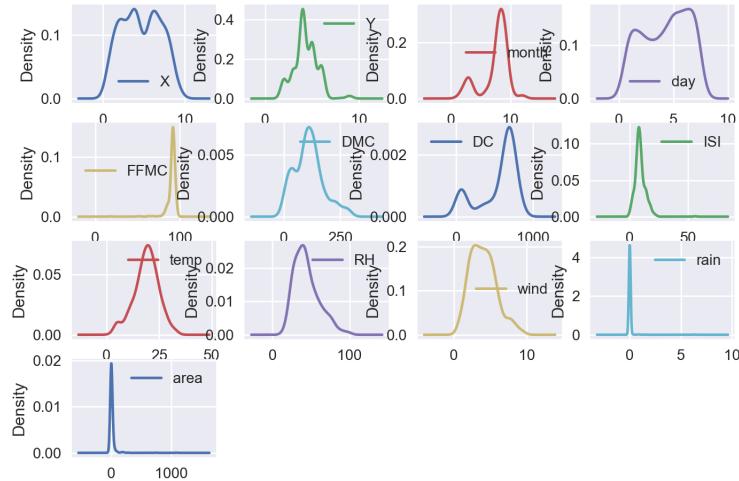
En nuestro caso, podemos ver como las variables DMC i DC son las que tienen una correlación más alta con un 0.68, seguido de las variables ISI i FFMC con un 0.53. En el primer caso suponemos que es un valor alto ya que la humedad en ambos casos debe ir relacionada, y si hay humedad en una capa, en la otra seguramente también haya.

En el segundo caso, al ser el índice de propagación inicial del sistema FWI, podemos entender que se trate de una correlación positiva. Por otro lado, podemos ver como muchas variables tienen una correlación prácticamente nula con la gran mayoría de las otras variables, es decir, que prácticamente no hay una correlación lineal. Por último, tenemos a la variable RH correlacionada con las variables temp y FFMC con los peores valores, con un -0.53 y un -0.3 respectivamente.

1.8 Conclusiones

En nuestro dataset está bastante claro que el atributo objetivo es el área, ya que el objetivo es ver como afectan los otros atributos a la cantidad de fuegos forestales para poder así controlarlos.

Una vez que tenemos claro el atributo objetivo, vamos a analizar los otros atributos para poder ver cuales serían más importantes a la hora de hacer una buena regresión. Para ello primero de todo vamos a ver que atributos siguen una distribución gaussiana. Para poder verlo mejor hemos hecho los siguientes gráficos:



En este gráfico se puede observar que los atributos que parecen que siguen una distribución gaussiana son los siguientes: ISI, temp, RH y wind.

Si tenemos en cuenta la tabla de correlación del apartado anterior, se puede ver que los atributos con mayor correlación con el atributo objetivo, el área, son la 'DMC', 'temp' y 'RH'. Dos de estas variables coinciden con las mencionadas anteriormente que siguen una distribución gaussiana.

Para poder tomar una decisión mas precisa sobre los atributos mas influyentes en nuestra regresión en el siguiente apartado vamos a hacer unos análisis usando datos estandarizados, diferentes regresiones...

2 Apartado B

2.1 Atributos

Hemos escogido los atributos DMC, ISI, temp y RH como los atributos más importantes. Hemos decidido esto después de ver los histogramas de los atributos

estandarizados, ya que estos tenían una distribución que se parecía más a una distribución gaussiana.

Hemos normalizado los datos y hemos podido observar que no varían mucho. Hemos utilizado tanto los datos sin estandarizar como los estandarizados para la regresión y no hemos podido ver mucha diferencia. Es más, de todos los modelos que hemos usado, sólo varían los resultados de Random Forest.

También hemos calculado los MSE (Mean Squared Error) de los atributos escogidos y hemos observado que los que obtienen un resultado menor son DMC y temp.

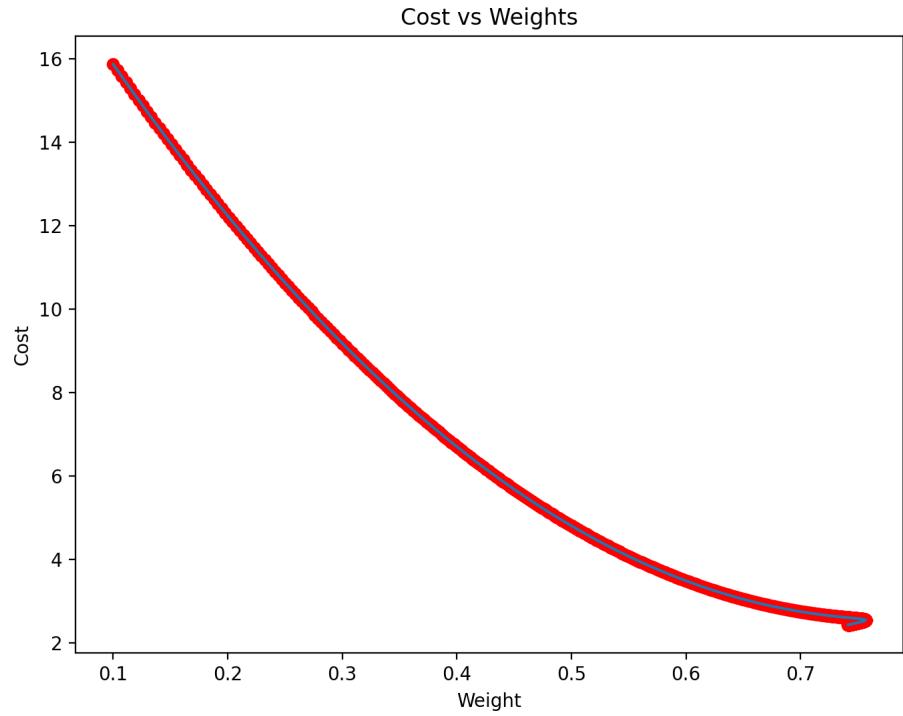
Ya que nuestro dataset tiene unos pocos outliers que son bastante extremos, hemos decidido quitarlos. Para ello hemos visto a partir de qué números de hectáreas de área quemada aparecen. Hemos visto que sólo hay 11 observaciones que estén entre los valores 100-1090.84 y hemos eliminado esos.

Esto ha ayudado con la regresión, ya que los resultados de MSE y R2 Score son mucho más bajos sin outliers.

2.2 Apartado A

El descenso del gradiente es un algoritmo que estima numéricamente dónde una función genera sus valores más bajos, es decir, que busca los mínimos locales. Para minimizar la función, el descenso del gradiente sigue el negativo del gradiente, y así irá en la dirección del descenso más pronunciado.

Para el descenso del gradiente necesitamos una función de coste, el objetivo es encontrar un conjunto de ponderaciones y sesgos que minimicen el costo. Una función común que se utiliza a menudo es el error cuadrático medio, que mide la diferencia entre el valor real de y y el valor estimado de y (la predicción). El resultado de esta función lo podemos ver graficado en la primera gráfica.



En la segunda gráfica tenemos en resultado de ejecutar la función del descenso del gradiente. Podemos ver que la gráfica está construida por un eje de 9x9, el cual representa el parque en el cual se producen los incendios, con puntos rojos donde se han producido los incendios y con la línea azul el resultado del descenso del gradiente.

