Universitat Autònoma de Barcelona

Aprenentatge Computacional

Pràctica 1 - Regressió

Autors:

Biel González NIU: 1551813

Judit Panisello NIU: 1605512

Cristina Soler NIU: 1603542

Octubre 2022

Índex

1 Introducció

Una de les principals preocupacions ambientals és l'aparició d'incendis forestals, que afecten la preservació dels boscos, creen danys econòmics i ecològics i causen patiment humà. Aquest fenomen es deu a múltiples causes com per exemple, negligència humana i llamps. I malgrat l'augment de les despeses estatals per controlar aquest desastre, cada any es destrueixen milions d'hectàrees forestals arreu del món.

La detecció ràpida és un element clau per a l'extinció d'incendis. Atès que la vigilància humana tradicional és cara i està afectada per factors subjectius, s'ha posat èmfasi en desenvolupar solucions automàtiques.

En contrast amb aquesta informació presentem un enfocament nou d'incendis forestals, on recopilarem dades recents del món real, que contenen informació valuosa, com ara tendències i patrons, que es pot utilitzar per millorar la presa de decisions en temps real i de forma no tan costosa.

Per tant, usarem eines de DM automatitzades per analitzar les dades en brut i extreure informació d'alt nivell per a qui pren les decisions.

2 Objectius

L'objectiu d'aquesta pràctica serà predir l'àrea cremada d'un terreny, és a dir, ser capaços maximitzar la confiança de prediccions, a partir d'un conjunt de dades recollides sobre diferents condicions ambientals i meteorològiques d'un mateix terreny, més concretament recopilades de la regió nord-est de Portugal.

Per a poder desenvolupar correctament aquest treball aplicarem les corresponents modificacions del dataset per a poder realitzar correctament els diferents tipus de regressions i així aconseguir unes bones prediccions.

3 Base de dades treballada - Fire Forest

El dataset seleccionat pel treball conté les dades d'incendis forestals en el parc natural de Montesinho, situat a la regió nord-est de Tr'as-os-Montes a Portugal. El parc té una gran flora i fauna, el clima és mediterrani i la temperatura mitjana anual està entre 8 i 12 graus Celsius. Les dades usades es van recollir entre gener del 2000 i desembre del 2003. El conjunt de dades va ser creat per Cortez i Morais (2007) fusionant manualment dos conjunts de dades. El primer va ser proporcionat per un inspector responsable de l'incendi de Montesinho, que va recollir dades temporals i espacials de cada incendi, juntament amb els components del sistema FWI i la superfície total cremada. El segon va ser recollit per l'Institut Politècnic de Bragança, i conté diverses observacions meteorològiques que van ser registrades en un període de trenta minuts per una estació meteorològica situada al centre del parc. Cada cop que es produïa un incendi s'enregistraven l'hora, data i localització espacial en una graella de 9x9 del parc, vegetació involucrada, els components del FWI, el vent, la pluja, la temperatura i l'àrea total cremada.

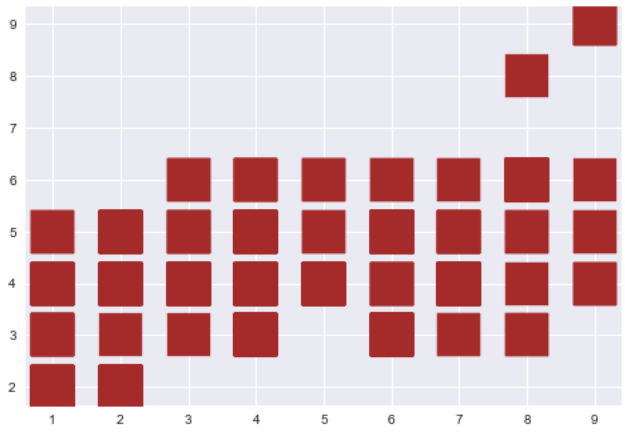


Figura 1: Representació simbòlica de l'àrea del parc

Per aquest estudi de dades s'ha utilitzat l'índex FWI (*Forest Fire Weather Index*). Creat per canadencs per trobar la manera de mesurar el perill dels incendis. Aquest índex està format per 6 components:

- FFMC : denota el contingut d'humitat de la brossa superficial i influeix en la ignició i propagació del foc.
- DMC: humitat en capes orgàniques poc profundes.
- DC: humitat en capes orgàniques molt profundes.
- ISI: puntuació correlacionada amb la propagació de la velocitat del foc.
- BUI: representa la quantitat de combustible disponible.
- FWI: indicador de la intensitat del foc, format a partir dels anteriors índexs.

3.1 Descripció de les variables

Finalment, obtenim un dataset amb 517 entrades i 13 atributs.

Atribut	Descripció
X	x-axis coordenada
Y	y-axis coordenada
month	mes de l'any (Gener a Desembre)
day	dia de la setmana (Dilluns a Diumenge)
FFMC	FFMC codi
DMC	DMC codi
DC	DC codi
ISI	ISI index
temp	Temperatura exterior (en C^o)
RH	Humitat relativa exterior (en %)
wind	Velocitat del vent (en km/h)
rain	Pluja (in mm/m^2)
area	Àrea cremada (in m^2)

Tabla 1: Atributs de la base de dades

3.1.1 Aclaracions

- X i Y: posició relativa a la graella del parc.
- FFMC, DMC, DC, ISI: són els quatre índexs bàsics del FWI que després formen la resta.
- rain: la pluja recollida trenta minuts abans de la declaració de l'incendi, la resta d'atributs son de mesures instantànies.
- area: els valors iguals a zero signifiquen que es va cremar una superfície inferior a $100m^2$.

4 Secció C - Analitzar Dades

L'objectiu d'aquest apartat és analitzar els diferents atributs que la componen la nostra base de dades, entendre'ls i fixar quin és l'atribut objectiu a predir de tots els que hi ha a la base de dades. Com observació, totes les gràfiques i taules es poden visualitzar amb millor resolució al notebook de jupiter.

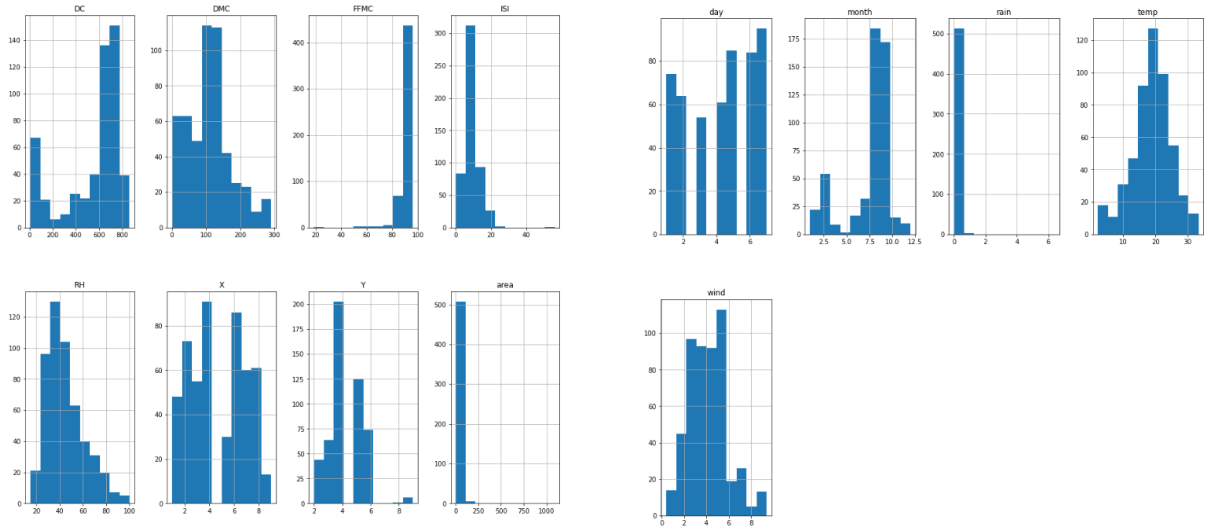
Primerament, hem mirat la dimensió del dataset d'on hem obtingut que està formada per 517 files i 13 columnes d'entrades. A més a més, s'han tractat els valors nuls i hem comprovat que no tenim cap dada nul·la.

Per visualitzar ràpidament les dades amb les quals tractarem hem fet un *df.describe()* amb les que podem visualitzar una descripció de les dades de cada columna.

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
count	517	517	517	517	517	517	517	517	517	517	517	517	517
mean	4.66	4.29	7.47	4.25	90.64	110.87	547.94	9.02	18.88	44.28	4.01	0.02	12.84
std	2.31	1.22	2.27	2.07	5.52	64.04	248.06	4.55	5.80	16.31	1.79	0.29	63.65
min	1.00	2.00	1.00	1.00	18.70	1.10	7.90	0	2.20	15.00	0.40	0	0
25%	3.00	4.00	7.000	2.00	90.20	68.60	437.7	6.50	15.50	33.00	2.70	0	0
50%	3.00	4.00	7.00	2.00	90.20	68.60	437.70	6.50	15.50	33.00	2.70	0	0
75%	7.00	5.00	9.00	6.00	92.90	142.40	713.90	10.80	22.80	53.00	4.90	0	6.57
max	9.00	9.00	12.00	7.00	96.20	291.30	860.60	56.10	33.30	100.00	9.40	6.40	1090.84

Tabla 2: Valors de les dades

Seguidament, hem fet un plot de les dades en forma d'histograma. D'aquesta manera hem pogut visualitzar la seva distribució i altrament comprovar quines dades tenen outliers. Com es pot apreciar a la figura 2. Aquí vam veure com les dades de: ISI, FFMC, area i rain contenien outliers, però no podien ser eliminats, ja que no són dades errònies si no valors vàlids fora del comú, per aquesta raó vam decidir mantenir-los durant l'anàlisi.



(a) (b)

Figura 2: Histogrames de les dades

Per realitzar un pairplot més concret vam afegir una variable a la taula, anomenada *Damage category* segons l'àrea cremada. Cada color representa una categoria diferent, on trobem: *No damage, high, low, moderate, very high*. Els vam assignar un valor seguint la següent mètrica:

Damage Category	**Area**
No damage	0
Low	≤ 1
Moderate	≤ 25
High	≤ 100
Very high	> 100

Finalment, tenim un gràfic amb les correlacions de les dades, podem veure que les dades del nostre dataset no tenen correlacions altes en valor absolut entre elles. És a dir tenim coeficients de correlació molt baixos i quasi nuls.

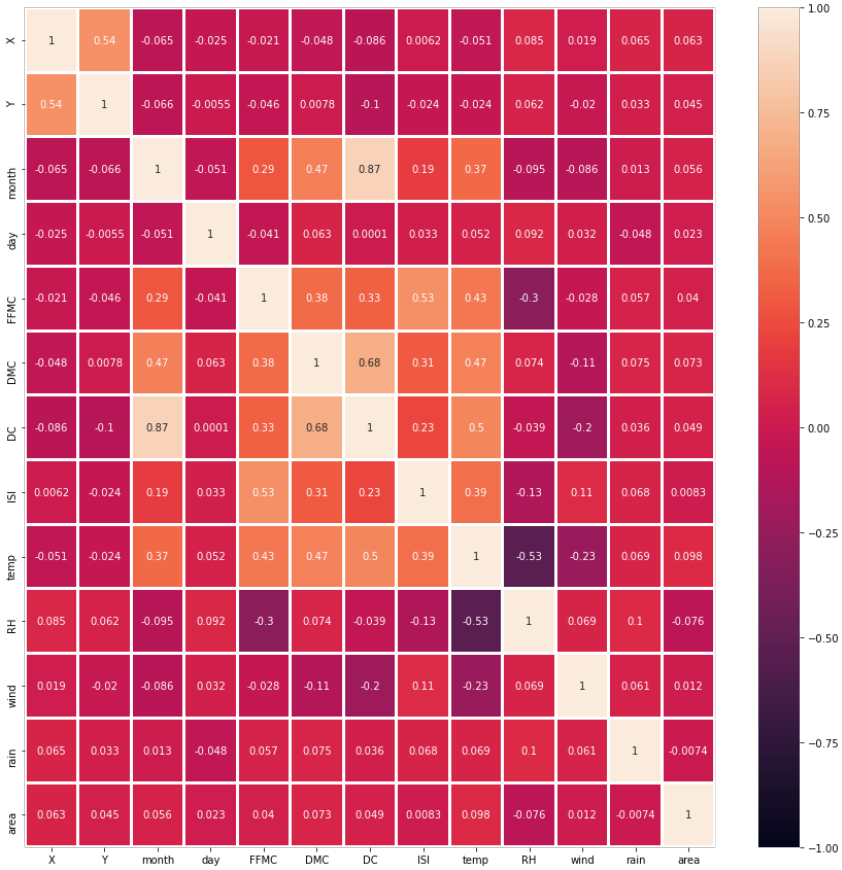


Figura 3: Heatmap de les correlacions de Pearson entre variables

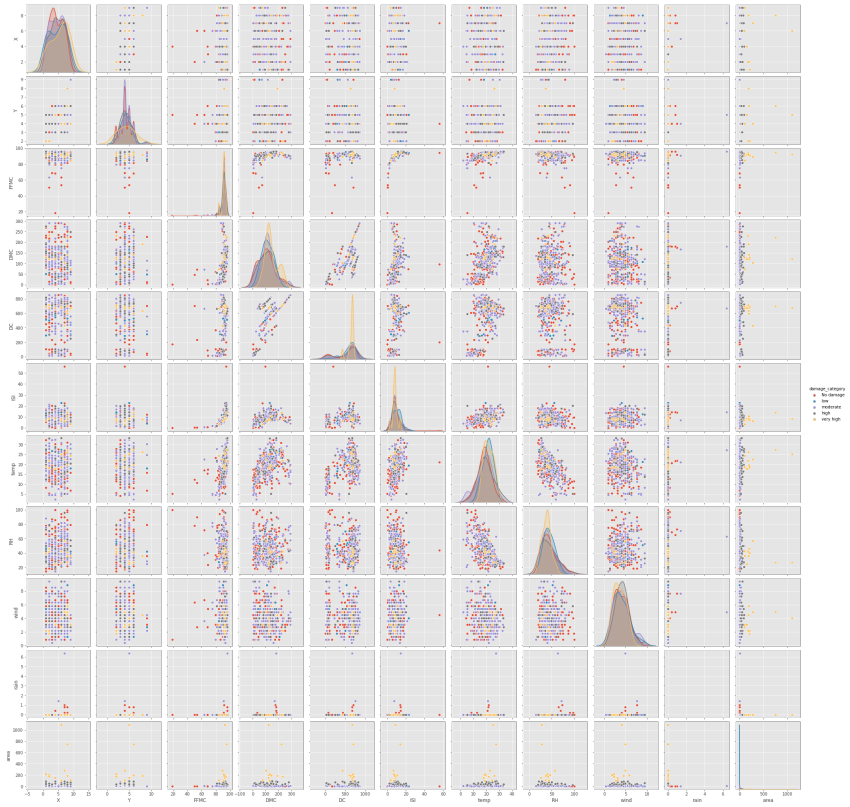


Figura 4: Pair plot de totes les variables combinades i colors en funció del dany produït

Les primeres dues columnes són les coordenades on s'han produït els incendis, estan indicades amb coordenades d'un mapa 9x9. En el següent gràfic es poden visualitzar les zones d'on tenim registre d'incendi. Com més marcat sigui el punt més incendis haurà patit la zona.

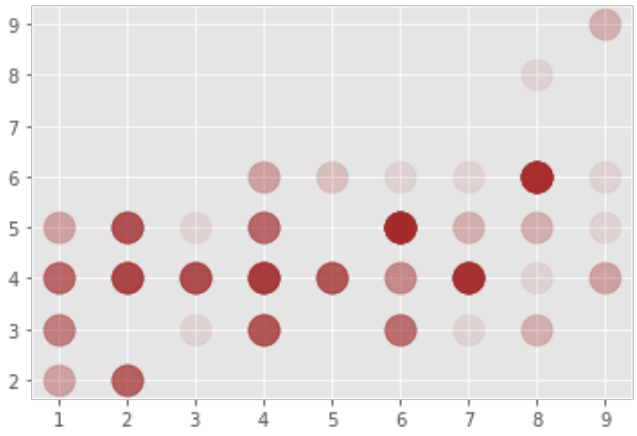


Figura 5: Incendis al parc

A partir d'aquest plot ens vam adonar que hi havia mesos on va haver-hi no només més incendis sinó que també de major extensió, així doncs vam mirar la freqüència dels incendis segons el dia o més de l'any per poder explorar amb major profunditat aquesta informació per conèixer millor el nostre dataset. Com podem veure a la figura 6 els mesos on s'han produït més incendis son agost i setembre, el qual és normal, ja que, normalment agost és el mes més calorós de l'any i setembre arrossega les conseqüències de la calor que hagi fet el mes anterior. Pel que fa al dia, diumenge és el dia amb major freqüència d'incendi, segurament perquè al ser cap de setmana més gent visita el parc, i el factor humà causa molts incendis; així i tot, no ho considerarem com a dada significativa, pel fet que ens faltaria més informació.

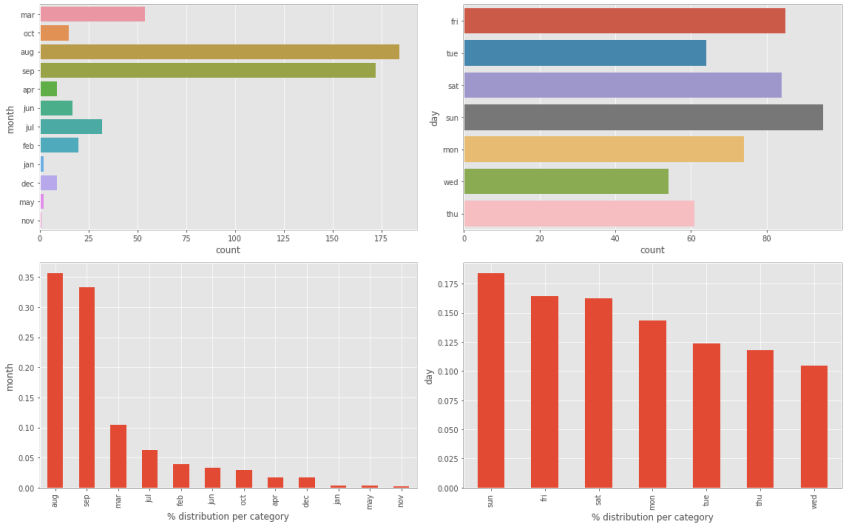


Figura 6: Freqüència d'incendis segons el mes i el dia de l'any

4.1 Preguntes a respondre

4.1.1 Clases de les variables

Variable	Tipus
X	int
Y	int
Month	int
Day	int
FFMC	float
DMC	float
DC	float
ISI	float
Temp	float
RH	int
Wind	float
Rain	float
Area	float

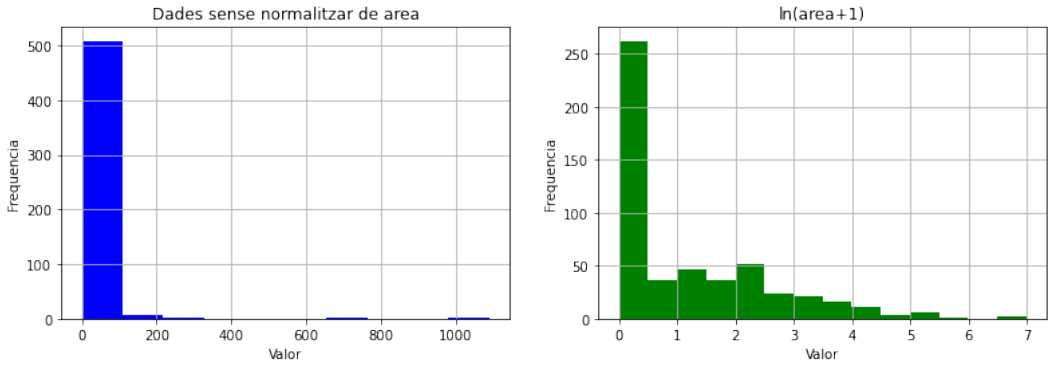
4.1.2 Possibles gaussianess

Després de fer un histograma de cada paràmetre podem veure com les dades dels atributs temperatura i l'ISI molt segurament segueixen una distribució normal o gaussiana. I la humitat relativa (RH) podria ser que també, però està desplaçada.

4.1.3 Atribut objectiu

Tal com ens indica el creador del dataset, la variable objectiu ha de ser l'àrea. Tot i això, per facilitar la interpretació i processament de les dades s'han de passar els valors per la transformació logarítmica $y = ln(x+1)$ i aconseguir així que estigui en una escala logarítmica. S'aplica aquesta transformació, ja que tendeix a millorar els resultats de regressió per a objectius inclinats a la dreta i la simetria. S'ha de tenir en compte que els nostres models utilitzaran l'atribut transformat en lloc de l'original. Hem triat aquest atribut, pel fet que creiem que és útil predir l'àrea que es cremarà en un incendi a partir de les dades ambientals.

Els següents gràfics mostren la diferència de les dades sense transformar i amb la transformació aplicada.



(a) Gràfica de l'area sense transformar (b) Gràfica de dades normalitzades

Figura 7: Efectes de la transformació

5 Secció B - Regressions

Ens hem trobat en diversos problemes per realitzar regressions, així que hem anat provant diferents tipus per provar quina regressió ens donava millors resultats.

5.1 Tractament de les dades

Abans d'efectuar cap mena de regressió sobre les nostres dades, durem a terme una normalització sobre aquestes per facilitar les tasques de regressió. Primerament, aplicarem la transformació logarítmica a l'àrea tal com s'ha indicat anteriorment.

Per algunes regressions com la lineal, multilineal, polinòmica i Lasso s'ha dividit el model en dues parts. Com hem vist anteriorment el valor de l'àrea té molts valors iguals a zero, això fa que el nostre model sigui complicat de modelar, ja que aquest zero no és un zero d'hectàrees cremades, sinó que l'àrea cremada en aquell incendi és inferior $0,01ha$ o també $100m^2$. Per tant, hem dividit el nostre model en dues parts. El primer model només conté els valors de l'àrea iguals a zero, en canvi, l'altre conté tots els altres valors diferents i majors que zero.

Altrament, separarem les dades en un train i un test, el train serà un 80% de les dades, i el test el 20% restant. Això ens dona un 216 de dades de Test. La normalització d'aquestes s'aplicarà a partir de la funció. *StandardScaler()*.

Dins del mateix Data Set hem trobat un conjunt de variables categòriques, les quals tenen una influència respecte a l'àrea cremada que volem predir, això ens porta al fet que hàgim d'aplicar un one-hot encoder sobre el dataset per a així poder tractar tot el conjunt d'atributs de manera conjunta en el moment de fer les diferents regressions.

5.2 Regressió lineal

5.2.1 Regressió lineal simple

El nostre primer intent de predir l'àrea cremada dels incendis forestals, en funció dels atributs disponibles, és mitjançant l'aplicació de la regressió lineal simple. És a dir la relació entre Y i X es modela com la següent combinació lineal:

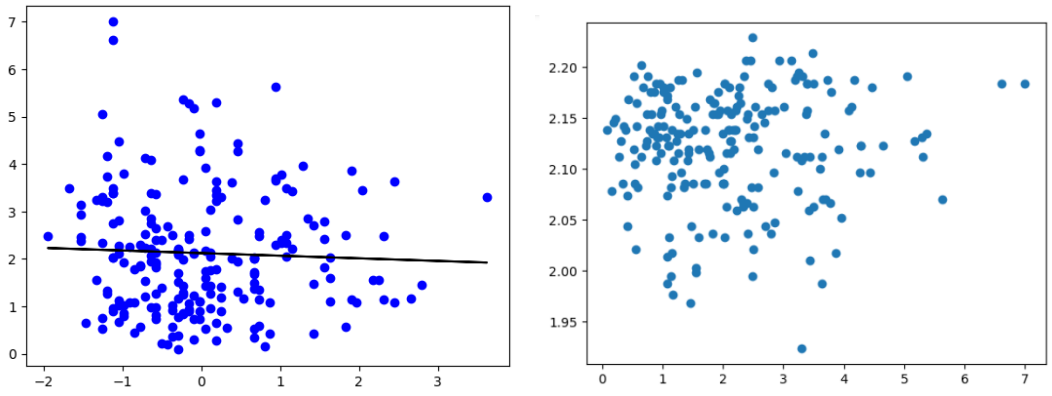
$$Y = X\beta + \epsilon$$

Per predir l'àrea cremada hem seleccionat l'atribut que més correlació té amb la variable de l'àrea. En el nostre cas la variable amb major relació de correlació amb l'àrea és *RH*.

Després d'entrenar el model obtenim: Com podem veure el coeficient de determinació és molt petit, per tant, no podem considerar que la variable *RH* com a única variable no és suficient per predir l'àrea cremada.

Coeficient de determinació	0.001849
Intercept	2.1202
Slope	-0.0549

Si analitzem els residus, també podrem veure que la regressió lineal no funciona. Per anar bé les dades de la gràfica del MSE haurien de quedar menys disperses, i aconseguir que les variables de y predita i y original fossin més semblants.



(a) Gràfica de la regressió lineal de RH (b) Gràfica de y-original i y-predita

Figura 8: Efectes de la transformació

5.2.2 Regressió lineal múltiple

Per continuar provant quin era el model de regressió bo, vam decidir provar amb la regressió lineal múltiple. Vam seleccionar dues variables que tinguessin molta correlació entre elles i també amb la variable objectiu. Les escollides varen ser *DC* i *DMC*, ja que entre elles tenen un elevat coeficient de correlació.

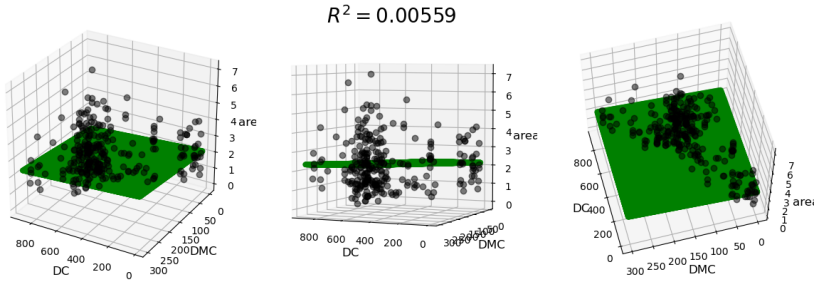


Figura 9: Regressió lineal múltiple

Com veiem el coeficient de determinació torna a ser molt baix, per tant, tampoc ens serveix.

Regressió agafant tots els atributs

Ja que en l'anterior regressió lineal múltiple no va funcionar com esperàvem vam decidir provar una altra regressió aplicant tots els atributs,siguin atributs categòrics o no. En aquest cas aconseguim el millor R score obtingut fins al moment, però com encara així no és prou bo per a predir correctament les dades de l'àrea cremada decidim continuar mirant altres tipus de regressions.

Coeficient de determinació	0.2061
MSE	1.2504
MAE	0.8620

5.3 Regressió polinomial

En veure que les regressions lineals no funcionen hem decidit provar amb la regressió polinomial. Aquesta ens permet visualitzar una relació no lineal entre les dades. La relació entre la variable X i la variable Y es modela amb un polinomi de n-éssim grau en X.

Per fer la regressió polinomial hem agafat el predictor de Temperatura. El qual augmentant fins a arribar al grau on r^2 fos màxim, obtenim que val 0.015, per tant, tornem a estar aproximadament a zero. És a dir la regressió polinomial tampoc ens serveix per trobar una predicció correcta de les dades. Com veiem a la següent imatge pel grau aconseguit la regressió comença ajustar-se a la zona de la dreta, però es manté en línia recta a la part esquerra.

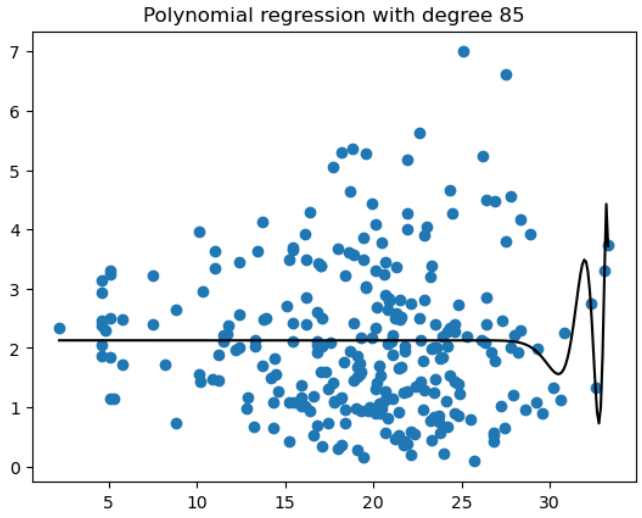


Figura 10: Regressió polinomial amb grau 85

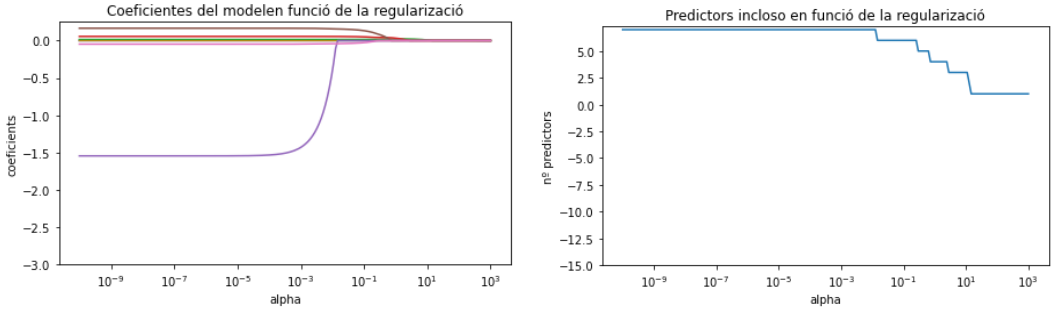
5.4 Altres regressions

5.4.1 Regressió Lasso

Donat que els resultats de les regressions anteriors vam decidir buscar un altre tipus de regressió que s'ajustés millor a les nostres dades. Per això vam escollir treballar amb la regressió Lasso. La regularització Lasso penalitza la suma del valor absolut dels coeficients de la regressió. Aquesta penalització força a fer que els coeficients dels predictors tendeixin a zero si no influeixen en el model. El grau de penalització està controlat per l'hiperparàmetre α. Quan $\alpha = 0$ el resultat és equivalent a un model lineal per mínims quadrats, a mesura que augmenta major és la penalització.

Per realitzar la regressió Lasso hem usat totes les variables predictores i hem avaluat quina és la millor α pel nostre model.

En fer servir una regularització, és útil avaluar com s'aproximen a zero els coeficients a mesura que α creix, així com l'evolució de l'error de validació. També podem veure com a mesura que augmenta el valor de α la regularització és major i més predictors queden exclosos.



(a) Coeficients del model en funció de la regularització

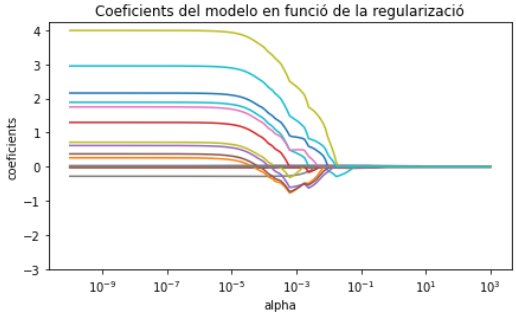
(b) Predictors en funció de la regularització

Figura 11: Anàlisis hiperparàmetre α

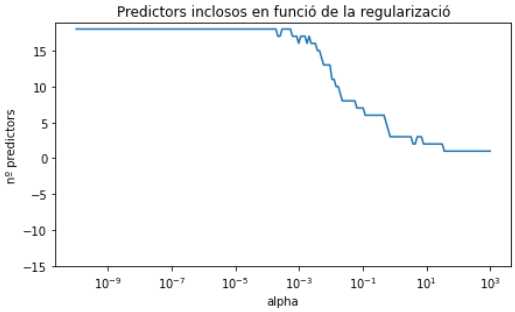
Trobem que la millor α pel model és 1e-10, per tant, l'assignem i fem la regressió. Com veiem es quedarà amb tots els coeficients, i obtindrem un MSE: 1.3617 i un r^2: -0.3868. Com que r^2 és negatiu indican una correlació negativa, en la que els valors d'una variable tendeixen a incrementar-se mentre que els valors d'un altre descendeixen. És a dir les variables es relacionen inversament.

5.4.2 Regressió Lasso i One-hot

Ja que l'one-hot amb mesos ens ha millorat una mica els resultats hem decidit combinar-ho amb la Regressió Lasso, per mirar si els resultats milloraven. Elaborarem la mateixa anàlisi que abans per trobar la millor α



(a) Coeficients del model en funció de la regularització



(b) Predictors en funció de la regularització

Figura 12: Anàlisis hiperparàmetre α

On obtindrem que el millor valor de α es mil. Això provocarà que el model penalitzi a tots els coeficients. Aconseguirem tots els coeficients amb valor 0, un MSE de 1.1567 i un r^2 de -0.00054, aproximadament zero, per tant, les variables no tenen relació entre elles. En sortir tots els coeficients amb 0 ja podem veure que no és vàlid.

5.4.3 Regressió logística

Com hem realitzat abans una columna a partir de l'àrea que ens feia una descripció de la gravetat dels danys al bosc, hem pensat a fer servir un concepte similar per saber si a partir d'una classificació senzilla sobre la gravetat de l'incendi si fer servir una regressió logística ens podria servir per poder predir si seria greu o no, així doncs aquest atribut substituiria l'àrea com a objectiu. Així doncs, tenim que les equivalències que hem assignat serien:

Damage Category	**Area**
Moderate (0)	≤ 25
High (1)	> 25

Trèiem l'àrea com a atribut a fer servir per a la regressió, ja que és una informació que està dins de la *Damage Category* de manera implícita i donar-li al regressor no tindria sentit.

A partir d'aquí realitzem una regressió logística amb ajuda de la llibreria skalearn. I fem un cross validation amb l'skalearn i obtenim d'aquí una accuracy d'un noranta per cent en tots els casos, el que a primera vista podria semblar molt bo. Així i tot, quan mirem millor els nostres resultats i fem una matriu de confusió per veure que tal ha anat, ens trobem que ho classifica tot com a incendis de dany *Moderate*, tots zeros, per tant, no podem dir que realment sigui una regressió efectiva o el que assolim com a resultat pugui ser útil, ja que pel que sembla no seria capaç de predir cap incendi de categoria *High*.

En veure això amb els atributs, vam decidir anar retirant atributs per veure com afectava això a la nostra regressió, i hem vist que igualment el resultat és el mateix, prediu només zeros, però cap u. Un cop vist això vam decidir canviar una mica la idea per la logística i buscar quins incendis serien classificats dins del zero, tots aquells que cremen menys de $100m^2$ i quins quedarien fora d'aquesta categoria. Però la precisió disminueix per

sota del cinquanta per cent i la matriu de confusió ens mostra que la classificació és molt dolenta.

5.4.4 Regressió logística + multivariada

Amb la idea de separar el nostre dataset en dues parts, una que contingués incendis més petits de $100m^2$ i una alta amb incendis més grans de $100m^2$, ens havia donat bons resultats amb la multivariada usant tots els atributs. Així doncs, si es pot classificar correctament en aquestes dues classes, podem obtenir una manera de predir a quina classe aniria i després depenent de la classe predir l'àrea cremada. Així i tot, en realitzar la

Damage Category	**Area**
Low (0)	$= 0$
Moderate(1)	> 0

regressió amb les dades obtenim una precisió pitjor a llençar una moneda a l'aire i decidir la classe a partir de si dona cara o creu. Per tant, tot i que els resultats traient els zeros a altres regressions siguin millors als que contenen zeros, no podem fer una classificació eficaç per continuar considerant-los dins del dataset i no ignorar-los.

5.5 Preguntes a respondre

5.5.1 Quins són els atributs més importants per fer una bona predicció?

Per la correlació de les dades amb la nostra variable objectiu no podríem extreure quines són les millors dades per realitzar una bona predicció, ja que totes tenen una correlació baixa i molt pròxima a zero. Amb la transformació de l'àrea tenim que les variables que tenen una correlació més alta són: *RH, DMC, ISI i wind.* Però com ja hem vist anteriorment és insuficient per dur a terme correctament una regressió.

5.5.2 Amb quin atribut s'assoleix un MSE menor?

Aplicant una regressió lineal obtenim els següents MSE, considerant que hem aplicat la transformació logarítmica a les dades, s'han eliminat els zeros i s'han normalitzat les dades.

Variable	**MSE**
X	1.6259
Y	1.6344
Month	1.6345
Day	1.6231
FFMC	1.6334
DMC	1.6328
DC	1.6288
ISI	1.6185
Temp	1.6239
RH	1.6315
Wind	1.6290
Rain	1.6301

Com podem veure tot i que les dades no són significatives, obtenim menys MSE amb el predictor *ISI*

5.5.3 Quina correlació hi ha entre els atributs de la vostra base de dades?

La correlació de les nostres dades com hem pogut veure abans 3 això ens mostra que les correlacions amb la nostra variable objectiu són molt baixes, tot i que a la vegada és aquesta la que té un major interès pel que fa a obtenir-ne prediccions. Això és congruent amb els resultats assolits, ja que de les regressions buscades i aplicades cap ha aconseguit treure resultats bons.

Altrament, podem fer la correlació amb les pertinents transformacions a la variable objectiu. Veurem que la correlació entre les dades no millora, encara és molt baixa.

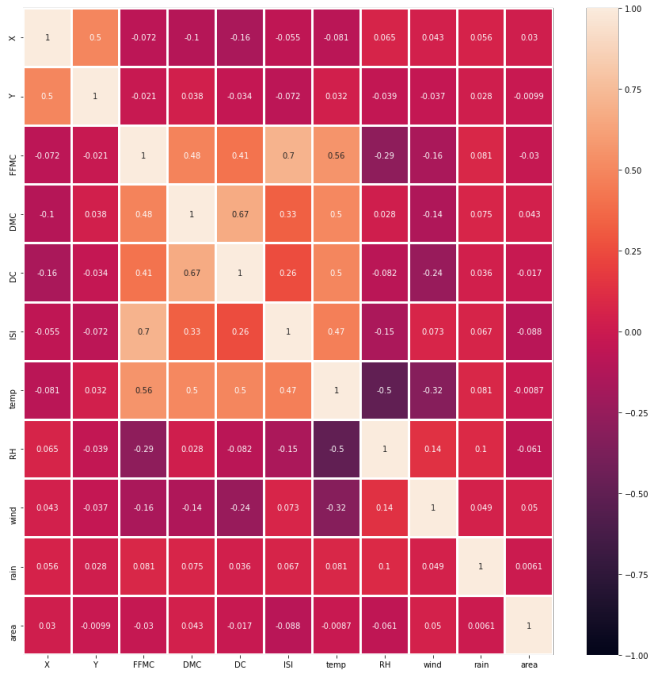


Figura 13: Correlació dades model sense zeros, amb transformació logarítmica a la variable objectiu

5.5.4 Com influeix la normalització en la regressió?

La regressió ens hauria de facilitar la interpretació de les dades, i fer que l'intercept sigui més significatiu. En el nostre cas no ha ajudat gaire, ja que les regressions han continuat essent no significatives.

5.5.5 Com millora la regressió quan es filtren aquells atributs de les mostres que no contenen informació?

En el nostre cas totes les mostres contenen informació així que no ha fet falta eliminar o emplenar columnes o files. Tal com hem comentat anteriorment vam decidir separar el model en dos. En un teníem tots els valors de zero a la columna d'àrea i al segon model els valors majors que zero. Això va implicar una petita millora de resultats, ja que a les regressions lineals vam passar de tenir un coeficient de determinació d'aproximadament zero a coeficients de determinació inversos, és a dir on les variables es relacionen inversament.

5.5.6 Si s'aplica un PCA, a quants components es redueix l'espai? Per què?

L'espai queda reduït de quaranta-quatre atributs a un total de quaranta-tres, és pel fet que un dels atributs després d'extraure les dades on l'àrea cremada que siguin iguals a zero, ha quedat sense informació que aportar i el PCA l'ha eliminat. Tot i que segons la gràfica de la PCA veiem que a partir de l'atribut trenta-sis apareix una línia recta la qual ens indica que no hi ha cap millora en el MSE ni en el R^2 i podríem reduir encara més el nombre d'atributs.

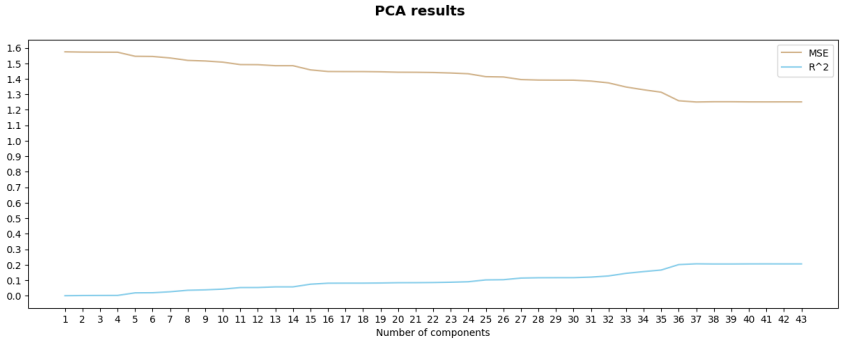


Figura 14: Gràfica del PCA

6 Conclusió

Hem provat molts tipus de regressions i amb diferents atributs i combinacions d'aquests i alterant una mica els objectius per poder realitzar una classificació però seguint el fil inicial del nostre propòsit. Hem vist que la millor regressió possible era la multivariada un cop havíem implementat l'One-hot a les variables dels mesos, dies, X i Y, i havíem prescindit de tots els incendis que havien cremat una àrea menor de $100m^2$. També hem provat mètodes de regressió més sofisticats com la Lasso per veure si obteníem uns resultats millors, però com hem vist, no ha sigut així i hem obtingut R^2 negatives el que segons hem vist significa que la predicció és pitjor que una línia que prediu el valor de la mitjana. Altrament, hem vist que si intentem fer una classificació els resultats no són satisfactoris i la classificació no és útil, ja que ho detecta tot com a incendis de tipus moderat. Així doncs, hem de decidir si la regressió amb l'One-hot i sumat a la pèrdua de la meitat del dataset és un augment prou significatiu per a donar-ho com bo.

D'acord amb el fet que tots els zeros que tenim no són zeros reals sinó omissions d'informació, com a grup hem arribat a la conclusió que la pèrdua de la meitat del dataset per a l'obtenció de la regressió multivariada és positiva i valida. Ja que hem obtingut la millor predicció de dades de totes les regressions i el núvol de punts de la predicció i el valor real és el que més s'acosta a una línia diagonal. Però s'ha de tenir en compte que la classificació per assolir si un incendi pertany a la classe dels menors de $100m^2$ o als majors de $100m^2$ dona resultats tan poc satisfactoris que seria simplement millor passar dels zeros com hem comentat abans.

Així doncs, concloem que la predicció de l'àrea cremada per incendis no és un problema tan simple que només depengui dels atributs que ens proporciona el dataset, ja que els fenòmens com els incendis són molt complicats de predir perquè hi ha molts més factors no només naturals. També cal comentar que els creadors dels datasets mencionen que les dades van ser obtingudes per dos grups diferents i després ajuntades. Probablement amb aquesta informació que falta es podria millorar la predicció i de l'abast dels incendis sota unes condicions donades.

7 Webgrafia

[1] https://www.cienciasinseso.com/transformacion-de-datos/

[2] https://www.kaggle.com/datasets/elikplim/forest-fires-data-set

[3] https://www.cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

[5] https://www.kaggle.com/code/psvishnu/forestfire-impact-prediction-stats-and-ml/notebook

[6] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[7] https://en.wikipedia.org/wiki/Principal_component_regression

[8] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[9] https://stackoverflow.com/questions/47442102/how-to-find-the-best-degree-of-polynomials

7.1 Github del projecte:

[10] https://github.com/Zynokrex/Regressio_APC