Pràctica 2: Com realitzar la neteja i l'anàlisi de dades?

Tipologia i cicle de vida de les dades
Gener 2022

Joan Peracaula Prat
Sara Jose Roig

1. Descripció del dataset

Per aquesta segona pràctica seguim utilitzant el mateix dataset que vam generar mitjançant web scraping en la primera part de la pràctica. En el nostre cas, vam construir un joc de dades de llocs on dormir en l'entorn natural d'Espanya, Andorra i Sud de França. Més concretament, els allotjaments recollits en el dataset són refugis de muntanya, tant lliures com guardats, i llocs d'acampada, tant càmpings com zones d'acampada lliure.

Com a recordatori, les dades van ser extretes del portal web https://www.walkaholic.me/ i publicades al lloc web de datasets Zenodo.

Tal com també es va argumentar en la pràctica 1, aquest conjunt de dades és important perquè recull unes dades que poden ser d'interès per gent aficionada a les excursions de muntanya, i també a organitzacions i institucions amb objectius com ara el mapatge d'aquests allotjaments o diverses anàlisis que en poden sortir derivats.

En el nostre cas, en aquesta pràctica ens plantegem respondre la següent pregunta:

Per aquest tipus d'allotjaments de muntanya (refugis i zones d'acampada),
quins són els atributs que més determinen si un allotjament és gratuït o pagament?

2. Integració i selecció

En aquesta pràctica, sobre el dataset original hem aplicat tant selecció, com integració de dades.

Selecció

Com és sabut, consisteix en seleccionar del conjunt de dades original aquelles variables que realment són interessants per a l'anàlisi posterior. Dins de l'etapa de neteja de dades, hem filtrat aquelles variables del conjunt de dades que tot i que puguin ser interessants per altres aplicacions, no ens aportaven valor al nostre anàlisi.

Per tant, hem filtrat columnes com ara la descripció de l'allotjament, el telèfon, web i email de contacte, entre d'altres. Les variables que hem seleccionat són les següents:

Variable	**Descripció**
Place type	Especifica quin tipus de lloc es tracta, refugi o espai d'acampada.
Name	Nom de l'allotjament. Tot i no ser important per l'anàlisi, permet identificar de forma única els allotjaments.
Place list	Regió geogràfica, en forma de llista de més a menys àrea. Segons el país, canvia l'estructura de la llista: - Espanya: país, comunitat autònoma, província - França: país, regió, departament - Andorra: país, parròquia
Capacity	Nombre de llits totals. (*Pot ser desconegut*)
Fee	Indica si és gratuït o de pagament.
Altitude	Altitud en metres sobre el nivell del mar en què es troba l'allotjament. (*Pot ser desconegut*)
Services	Serveis que ofereix l'allotjament en forma de llista. Inclouen: bany, calefacció, dutxa, internet, llar de foc, mantes, menjar i begudes, ràdio d'emergència… (*Poden ser desconeguts*)
Coordinates	Coordenades de latitud i longitud de la localització de l'allotjament.
Nearby routes	Llista de noms de rutes de senderisme que es troben a prop de l'allotjament. (*Poden ser desconegudes*)

Integració

Consisteix en combinar dades procedents de fonts diverses. En el nostre cas, el dataset conté dades força completes per la majoria dels atributs importants seleccionats, però per alguns d'ells tenim valors buits o desconeguts. En l'apartat de neteja de les dades entrarem més en detall, però com a avançament, els atributs amb més valors zeros o buits eren *Capacity, Altitude, Services i Nearby routes*.

Vam plantejar afegir un procés d'integració de dades amb l'objectiu de corregir valors buits de les variables seleccionades. D'aquestes quatre variables, n'hi ha tres que no era viable obtenir

aquestes dades per altres fonts (*Capacity, Services i Nearby routes*), però per la variable Altitude sí que vam trobar una manera viable de completar les dades restants: mitjançant les coordenades de latitud i longitud i una font externa en forma de DEM (*Digital Elevation Model*).

Així doncs, la fase d'integració de dades ha consistit en, primer, obtenir un DEM de la regió geogràfica de les dades del dataset (Espanya, Andorra i sud de França) i després utilitzar-lo per extreure una elevació (aproximada, però bastant precisa) en funció de les coordenades. L'etapa d'obtenció d'un DEM d'aquesta regió concreta no ha estat fàcil, ja que els DEMs *open-source* es descarreguen per regions petites, i ha calgut descarregar-los separadament i combinar-los tipus mosaic en un sol fitxer DEM (el codi per aquesta tasca també es pot trobar en el repositori).

Tot i que aquí en la memòria aquesta etapa apareix prèviament a l'apartat de neteja de dades, en el nostre context, ens era més adequat fer-ho després de netejar i preprocessar les coordenades i l'altitud original de les dades del dataset.

Per últim, comentar que aquest procés d'integració ha permès passar de 581 registres amb altitud desconeguda inicialment a només 12, que pels motius que sigui (imprecisió, forats de dades o fora rang), el DEM n'ha retornat valors extrems que hem marcat com a valors Null.

3. Neteja de les dades

Després de la selecció i abans de la integració de dades, hem realitzat un procés de neteja de dades per poder-les utilitzar en un posterior anàlisi. Aquesta neteja ha consistit en diversos processos. A la taula de continuació hi ha especificat per variable quin tipus de canvis s'han realitzat:

Variable	**Canvi realitzat**
Place type	Canvi de nom de l'atribut a minúscules i sense espais: place_type. Transformació de variable binària amb etiquetes a variable binària amb valors: 0 si el registre es tracta d'un refugi o 1 si es tracta d'una zona d'acampada.
Name	Canvi de nom de l'atribut a minúscules
Place list	Segmentar la variable en 3 diferents categories: country, region i place. Hem eliminat les paraules caràcters sobrants per tenir només el nom del lloc. En el notebook d'anàlisi, codificació d'etiquetes a valors numèrics de tipus integer.
Capacity	Canvi de nom a minúscules i canvi de tipus de variable a integer. Valors desconeguts ('? beds') marcats com a NA, ja que no podem afirmar que

	siguin valors 0.
Fee	Canvi de nom a is_free i transformació a variable binària, on 0 indica que l'allotjament és de pagament i 1 que és gratuït.
Altitude	Transformació a variable numèrica, substituint els valors desconeguts ('?') per NA, ja que tampoc és correcte posar altitud 0m.
Services	Canvi a variable anomenada num_services, on en lloc de tenir una llista de serveis, és un integer que és la suma del número serveis per allotjament. A causa del nombre elevat de registres amb llista de serveis buida, creiem que és degut a desconeixement i no a valor zero, per això assignem NA a aquests casos.
Coordinates	Extreure les paraules sobrants, i segmentar la variable en dues variables separades latitud (latitude) i longitud (longitude). Canvi a variables tipus float.
Nearby routes	Procés similar a la variable Services, canvi a variable anomenada num_nearby_routes on aquesta és el nombre de rutes que hi ha pròximes a l'allotjament. En aquest cas sí que associem a llista buida com a valor 0.

Tal com s'ha explicat en la taula, els valors buits o desconeguts s'han gestionat de forma diferent segons el context de cada variable. En general, hem tractat els valors desconeguts assignant-los a valors NA, ja que creiem que és el valor més adequat per aquestes variables en el dataset final. Tot i això, en l'anàlisi, per alguns casos s'han deixat com a NA i pels models que no accepten valors buits, s'han eliminat aquestes entrades i només s'han avaluat la resta de valors.

4. Anàlisi de les dades

Com ja hem mencionat al primer apartat, l'anàlisi de dades l'hem centrat en determinar quins són els factors que més influeixen en si un allotjament és gratuït o de pagament. Per arribar a una conclusió hem elaborat diferents anàlisis: univariant, bivariant, multivariant, correlació, comprovació de normalitat i regressió logística.

A continuació detallem els anàlisis aplicats juntament amb les més destacades representacions dels resultats. El codi d'aquests anàlisis es pot trobar en el jupyter notebook anomenat *'data_analysis.ipynb'*.

Anàlisi univariant

L'anàlisi univariant s'ha realitzat a partir d'histogrames i d'un anàlisi descriptiu de les dades amb valors mínim i màxims, la mitjana i la derivació estàndard com a dades importants. Amb aquesta

informació ens podem fer una idea de les característiques dels allotjaments del joc de dades i els valors de les variables.

Com podem veure, la majoria d'ells són refugis i hi ha una proporció més petita de llocs d'acampada. El nombre de llits disponibles oscil·la majoritàriament entre 0 i 50 amb una mitjana de més de 19 llits per allotjament i la majoria són de pagament. Com era d'esperar la latitud i longitud oscil·la dintre un mateix rang que és el de les coordinades d'Espanya, França i Andorra; i l'altitud és força variable amb una mitjana de 1146m. Pel que fa al nombre de serveis, aquest oscil·la entre 1 i 12. I el nombre de rutes properes entre 0 i 97, on la majoria d'allotjaments en tenen menys de 30.

Anàlisi bivariant

A continuació, a la figura 1 podem veure una gràfica on comparem el tipus de lloc 0 per refugi i 1 per lloc d'acampada amb si és de pagament o no. La primera conclusió que podem treure és que tots llocs d'acampada són gratuïts i la majoria de refugis sí que ho són de pagament.

A continuació podem veure una sèrie de gràfiques tipus boxplot. En aquests gràfics podem veure on el gruix més gran de valors, els valors extrems i la mediana de les variables numèriques comparades amb la variable is_free. Alguns d'ells ens donen conclusions que ja eren les esperades, com per exemple que la latitud, longitud i altura no influeixen en que un lloc sigui de pagament o no. En canvi, en les gràfiques del nombre de rutes properes i el nombre de serveis de l'allotjament, donen uns resultats contraris als esperats, els allotjaments amb més nombre de serveis i més nombre de rutes són els gratuïts.

A la figura 5 podem veure unes gràfiques tipus scaterplot on hi ha comparacions d'entre totes les variables. És útil per visualitzar totes les variables alhora, això no obstant, per la distribució de les dades no podem veure cap correlació lineal ni treure cap conclusió nova.

Anàlisi multivariant i de correlació

Per extreure més informació de les variables sobre la seva correlació entre elles de la que hem pogut veure en els scatter plots, hem fet un anàlisi multivariant. A la figura 6 podem veure un heatmap que indica la correlació entre variables, a major correlació, és a dir, quan és més propera a 1, més intens és el color. Si obviem les comparacions d'una variable amb ella mateixa on el resultat és evidentment 1, podem veure que les variables que tenen més correlació entre elles són nombre de serveis amb capacitat, capacitat amb tipus de lloc i latitud amb nombre de rutes properes. I les que menys, i doncs, amb més correlació negativa, són la variable is_free amb nombre de serveis i capacitat.

A continuació hem comprovat el coeficient de correlació amb la variable target is_free. Amb el coeficient de Pearson podem veure la força de la relació lineal de dues variables. El valor oscil·la

entre 1 i -1, on el 0 indica que no hi ha cap tipus de correlació, i on valors per sobre de 0,5 o -0,5 indiquen una correlació notable. En el nostre cas podem veure que la variable is_free té una correlació notable amb les variables num_services tal com ja havíem vist també en el heatmap.

Normalitat i Homogeneïtat

A la gràfica 3 podem veure la distribució que només la variable altitud té una distribució normal i que les altres variables numèriques presenten distribucions no normals. El coeficient d'asimetria confirma aquesta mateixa informació. Quan el valor d'aquest coeficient és proper a 0, la variable presenta una distribució normal, si és més gran que 1, la distribució és més densa cap a l'esquerra. En cas que sigui menor a -1 la distribució seria més densa cap a la dreta. No obstant això, valors més grans i més petits de 0.5 i -0.5 ja tenen un 'skewed' moderat. En el nostre cas latitud i longitud serien més denses cap a la dreta, num_nearby_routes seria moderat.

Per poder entrenar aquestes dades en diferents models és aconsellable arreglar aquestes distribucions perquè siguin més normals. Per aconseguir-ho hem aplicat una transformació logarítmica. A la figura 4 podem veure el resultat després d'aquesta transformació.

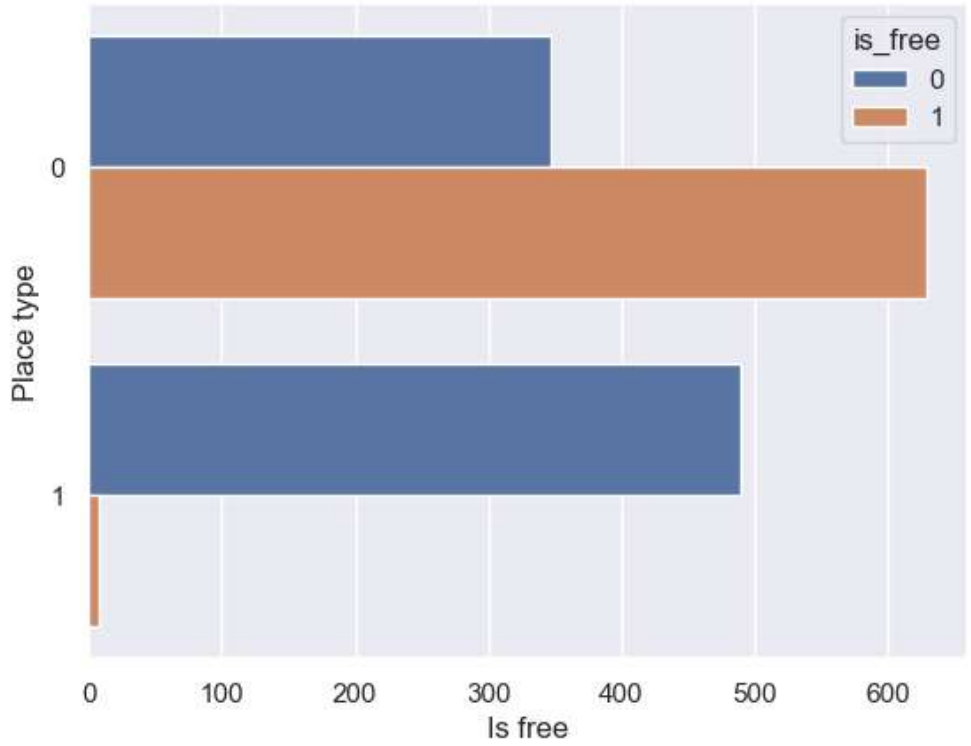
Regressió logística

Per poder fer una classificació entre allotjament gratuït i de pagament hem aplicat un model de regressió logística. Primerament, hem partit les dades entre un grup per entrenar i un grup per testejar, amb una partició de 0,7/0,3. A continuació, hem entrenat el model i hem fet les prediccions. Per comprovar el rendiment del model hem creat una matriu de confusió on podem veure que els veritables positius i negatius són uns valors molt més alt que els falsos positius i negatius, cosa que indica un bon rendiment del model. També podem veure que l'exactitud, és a dir, el nombre de dades classificades correctament; la precisió, com és de bo el model per assignar verdaders positius com a tals; i la f1-score, que és una combinació de la precisió i la sensibilitat, presenten valors alts.

5. Representació dels resultats

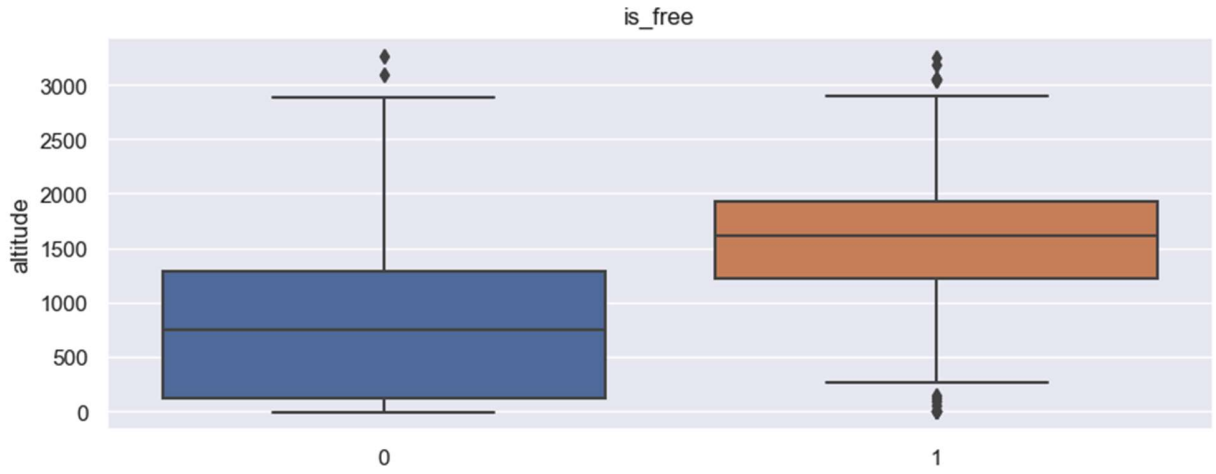
Les representacions dels resultats es poden trobar en el notebook d'anàlisi juntament amb algunes explicacions, però a tall de resum aquí es mostren les representacions més rellevants.

En primer lloc, mostrem la distribució de valors de les dades comparant entre tipus d'allotjament i si és gratuït. *Place type* 0 fa referència a refugis i 1 a zones d'acampada. *Is free* = 1 significa gratuït, i 0 de pagament.



Podem veure que, comparativament, les dades prenen valors molt diferents segons aquests dos parells de variables. Per una banda, veiem que el nombre de refugis és més elevat que el nombre de zones d'acampada, però a la vegada veiem que la gratuïtat o no predomina contràriament segons el tipus. Això ens dona una pista gran que la variable tipus serà segurament determinant.

Continuem amb un segon gràfic comparatiu. En aquest cas de tipus boxplot, on mostrem la distribució de valors de la variable altitud partint en dos grups: allotjaments de pagament i lliures. Podem veure que els quantils centrals pels dos grups estan força diferenciats, indicant que la variable altitud pot ser una bona candidata per ajudar a determinar si és de pagament o no.

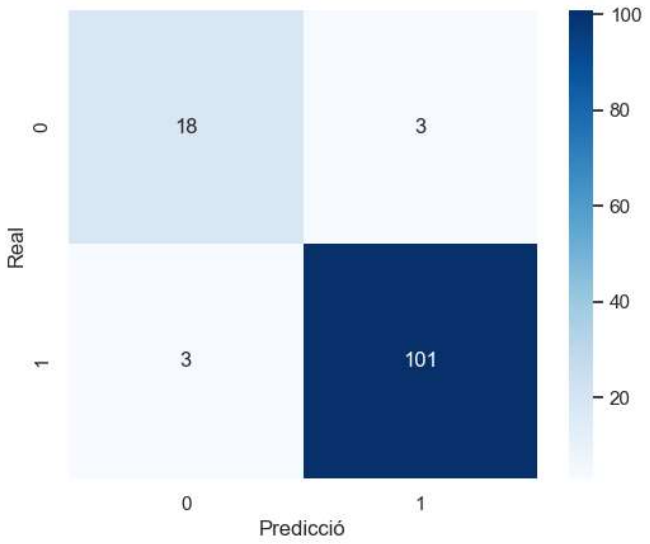


Les altres variables que han mostrat una correlació forta amb la variable de gratuïtat, són la capacitat i el nombre de serveis. A continuació es mostren les correlacions de Pearson, remarcant les més fortes, positivament o negativament:

```
Correlation is_free with place_type: -0.6014766567738076
Correlation is_free with capacity: -0.4763317815691487
Correlation is_free with altitude: 0.5109127490529971
Correlation is_free with country_enc: -0.21887384662303888
Correlation is_free with region_enc: -0.07686276094814401
Correlation is_free with num_nearby_routes: 0.12321871944793428
Correlation is_free with num_services: -0.3436031955577529
```

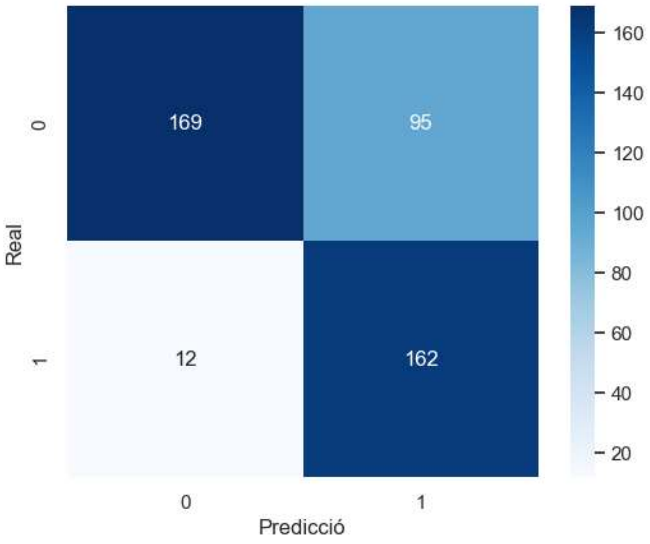
D'aquestes variables amb correlacions més fortes, ja hem parlat del tipus i l'altitud, que justament mitjançant inspecció visual en gràfics concrets, hem pogut esperar correctament que mostressin una correlació alta.

Ara ens queden la capacitat i el nombre de serveis. Aquestes dues tenen correlacions una mica més baixes, però igualment destacables, sobretot la capacitat. Ja hem comentat en l'apartat d'anàlisi hem aplicat una regressió logística, de fet han estat dues. En primer lloc, hem agafat com a variables dependents aquestes quatre variables. Per tal de poder aplicar un model de regressió logística calia prescindir de les files amb valors NA, que justament hem vist que les variables capacitat i nombre de serveis tenien força valors desconeguts. Això ha provocat que el nombre de dades per entrenar i testejar el model quedés molt reduït. En l'avaluació del model hem detectat una falsa quasi perfecció en la precisió, a causa d'un mal ajustament del model per la poca quantitat de dades i el desequilibri d'elles respecte a la variable objectiu.



Matriu de confusió del primer model, podem veure que hi ha força biaix en les prediccions

Així que hem decidit entrenar un segon model descartant aquest parell de variables amb tants valors NA. El segon model, ara sí, tenia un nombre raonable de dades ben equilibrades respecte a la variable objectiu. Després d'entrenar i testejar aquest segon model hem vist que tot i obtenir una precisió inferior (75%), aquesta segueix essent una bona mètrica i sobretot les prediccions no han sortit esbiaixades, a diferència del model anterior.



Matriu de confusió del segon model

En el notebook d'anàlisis es poden veure tots aquests detalls amb les seves explicacions. A continuació passem a les conclusions finals i resposta de la pregunta plantejada inicialment.

6. Resolució del problema

Com a conclusió d'aquest anàlisi, hem vist que **les variables que més determinen si un allotjament d'aquest conjunt de dades és lliure o de pagament són: el *tipus*, refugi o zona d'acampada, juntament amb *l'altitud***.

Hem vist que les variables *capacitat* i *nombre de serveis* tenen certa correlació amb la variable de gratuïtat, però a causa de la poca quantitat de valors coneguts per aquestes variables, no podem considerar-les com a suficientment determinants.

Per últim, les variables que no influeixen en si l'allotjament és de pagament o no són la *latitud*, la *longitud*, el *país*, la *regió* i el *nombre de rutes* que hi passen.

7. Codi

El codi de la pràctica es pot trobar en el següent repositori de Github:
https://github.com/sarajose/data_analysis_of_a_dataset

8. Vídeo

Link del vídeo explicatiu de la pràctica:
https://drive.google.com/file/d/17Ik3bn7iIUwxXC8qljhpBuQI5Gfvk5VI/view

9. Contribucions

Contribucions	**Signatura**
Investigació prèvia	SJR, JPP
Redacció de les respostes	SJR, JPP
Desenvolupament del codi	SJR, JPP
Participació al vídeo	SJR, JPP