Aprenentatge automàtic

Documentació del projecte - Abalone age prediction

Alex Iniesta i Adriàn Lozano

Quatrimestre de tardor 2019-2020

APA

Índex

1. Introducció

En aquesta pràctica se'ns proposa desenvolupar un model de classificació o de regressió per a resoldre un problema de tria pròpia.

D'entre els problemes proposats en la guía de la pràctica em escollit el que tracta sobre els abalones, les seves dades es poden obtenir a:

https://archive.ics.uci.edu/ml/datasets/Abalone

Els algorismes, anàlisi i preprocessat que durem a terme sobre les nostres dades el realitzarem amb el llenguatge de programació Python, hem triat aquest llenguatge per la nostra familiaritat amb el mateix.

D'entre les llibreries disponibles per afrontar el problema, nosaltres hem fet ús, principalment, de Numpy, SKLearn (per la creació de models), Pandas (per la lectura de les dades), Matplotlib, Seaborn i Scipy.

El nostre conjunt de dades, *The abalone dataset,* va ser originalment publicat per la UCI, el problema original, i el que nosaltres intentarem resoldre, és estimar la edat dels *abalone* (orelles de mar). L'edat dels *abalone* pot ser determinada a base de contar el número d'anelles que té a la closca, això però, és un mètode costós i feixuc, per tant la idea es aproximar una solució amb mesures més fàcils de fer, com són l'altura, el diàmetre o el pes. El que farem llavors és resoldre un problema de regressió lineal amb aquestes dades per predir l'edat dels *abalone*.

2. Anàlisi i preprocessat de les dades

2.1 Anàlisi

En aquest apartat farem un anàlisi previ de les nostres dades, en farem un anàlisi univariat i multivariat.

Anàlisi univariat

Comencem per l'atribut objectiu, podem observar que els seus atributs varien entre 1 i 29 anells, de totes maneres els valors més freqüents els trobem al voltant de la mitjana de la distribució. Veiem també que la distribució segueix una normal.

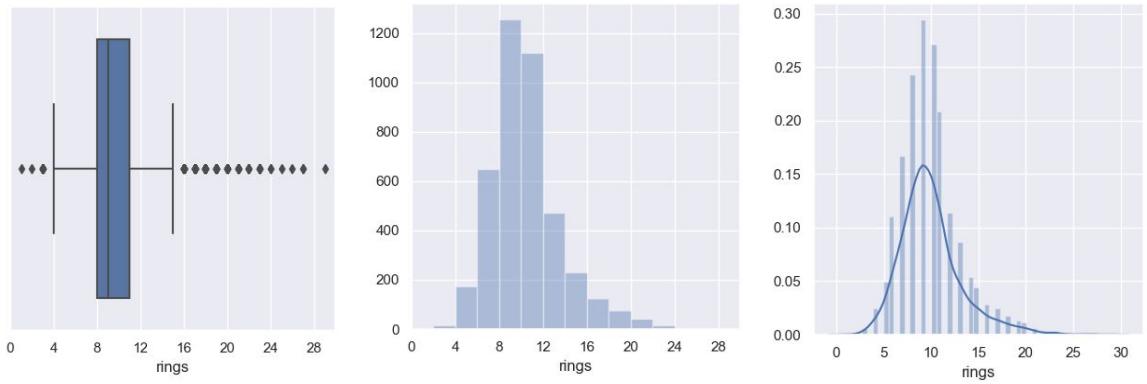

Figura 1. Boxplot, distribució i distribució normalitzada de la variable objectiu

Seguim amb els atributs de mida, llargada, diàmetre i altura. Podem observar com, igual que en el cas anterior, la distribució de les mesures s'aproxima a la normal, tot i així observem com en el cas de l'altura tenim un pic molt elevat degut a dos observacions molt llunyanes a les posicions centrals de la distribució.

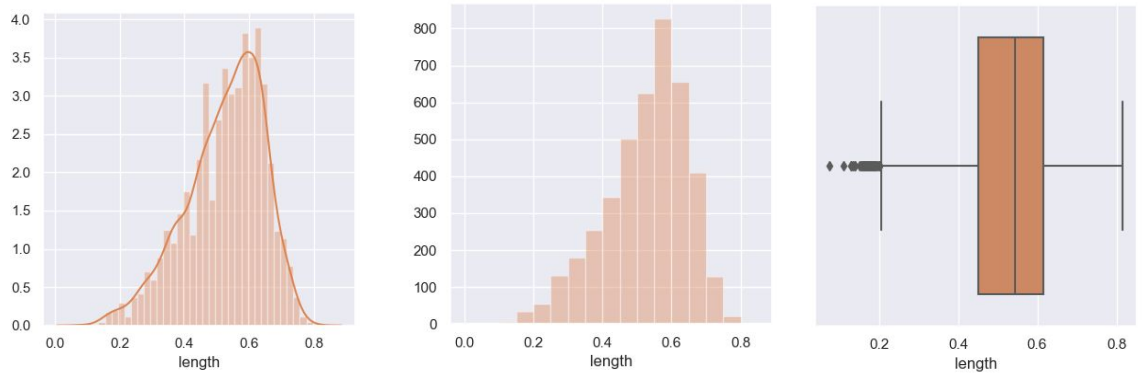

Figura 2. Distribució normalitzada, distribució i boxplot de la variable longitud

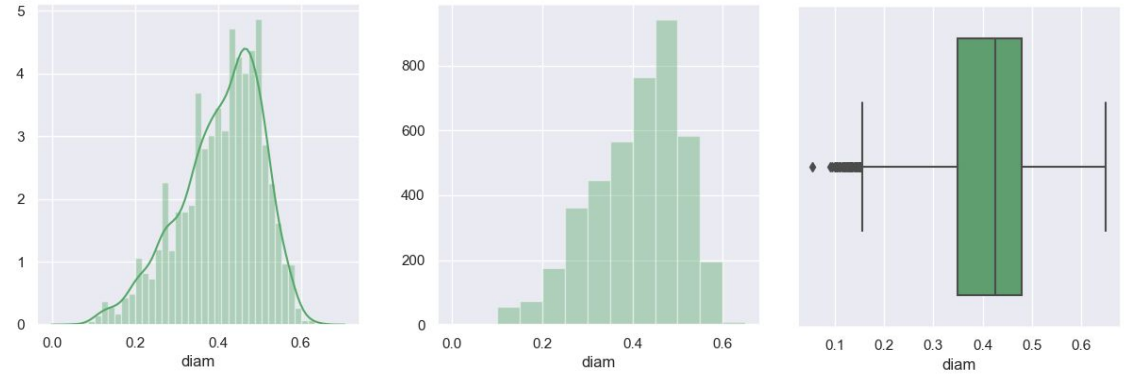

Figura 3. Distribució normalitzada, distribució i boxplot de la variable diàmetre

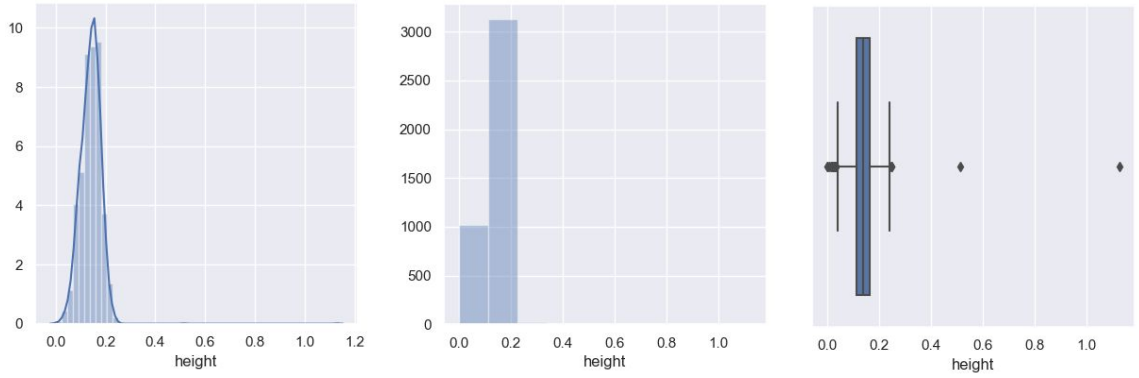

Figura 4. Distribució normalitzada, distribució i boxplot de la variable altura

Si filtrem aquests outliers obtenim una representació més realista.

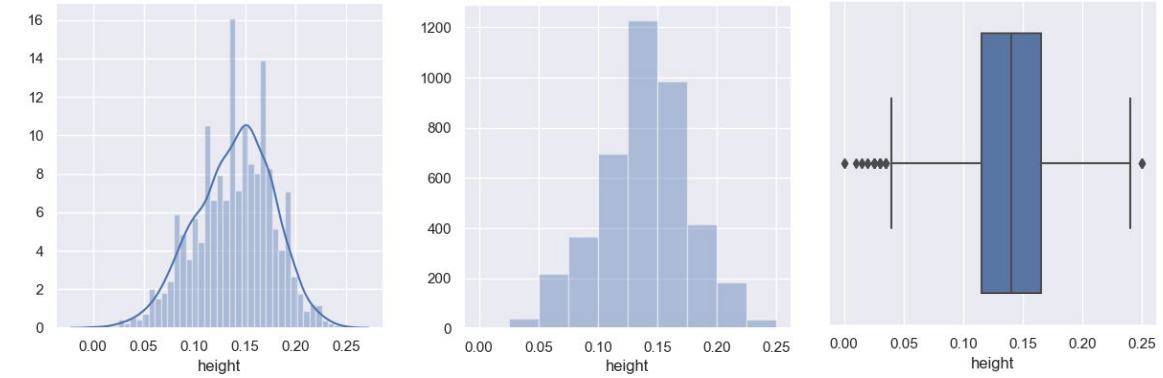

Figura 5. Distribució normalitzada, distribució i boxplot de la variable altura sense els outliers

Anàlisi multivariat

En aquesta part del anàlisi observarem la correlació entre les diferents variables, per veure com es relacionen amb la variable objectiu i entre si. Analitzant la matriu de correlació observem que els atributs de l'altura i el pes de la closca son els que millor correlació tenen amb la variable objectiu. La correlació de les variables de sexe les podem obviar.

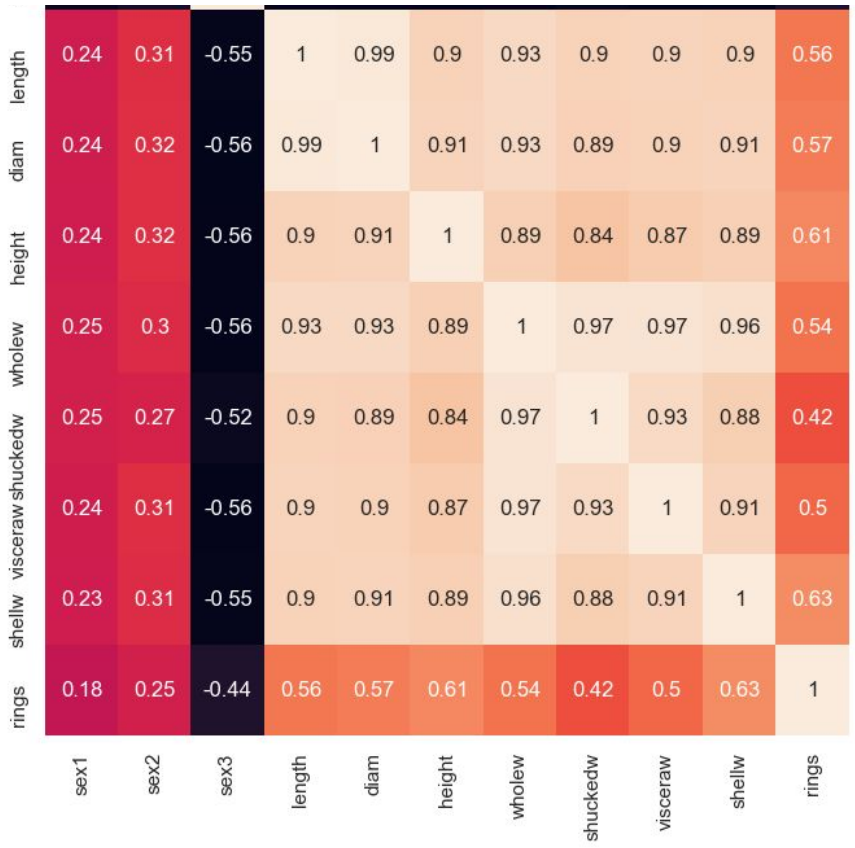

Figura 6. Matriu de correlació de les variables

La següent observació es veure si la correlació canviava amb el número d'anells, degut al que havíem llegit sobre la mostra sospitavem que la relació entre les variables canviaria de joves a adults, com podem veure en les següents matrius de correlació, que corresponen a les dades que tenen menys de deu anells i als que tenen 10 o més respectivament, les dades amb anells menors a 10 tenen una correlació significativament millor.

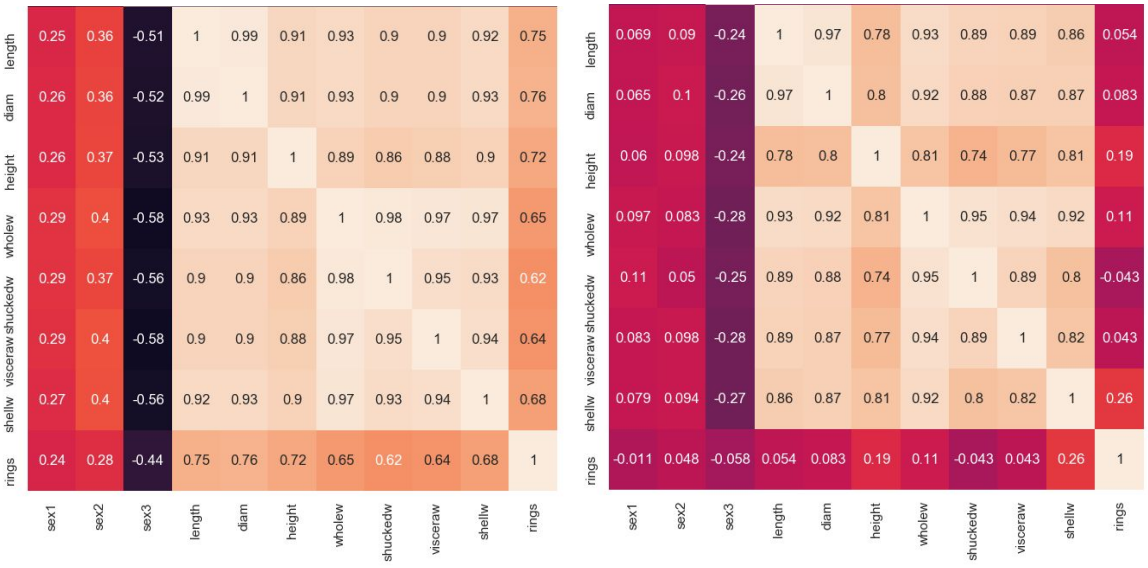

Figura 7. Matriu de correlació de les variables amb menys de deu anells i amb més de deu anells

2.2 Preprocessat

Com hem vist en el nostre anàlisi i degut al tipus de les nostres dades no tenim moltes coses que fer en el processat, per tant l'únic que li aplicarem a les dades es un filtre per la seva *z score* de tal manera que poguem detectar els outliers, aquests outliers detectats els eliminarem ja que com hem pogut veure en l'anàlisi aquests seran probablement, errors en la mesura.

3. Mètodes de resolució i resultats obtinguts

Per resoldre el problema presentat a l'*abalone dataset*, com hem mencionat, hem optat per tractar-ho com un problema de regressió, encara que es podia haver plantejat com un de classificació. Aquesta primera decisió és deguda a que hem considerat que modelitzar l'edat com un valor continu en comptes d'un valor discret tindria més sentit a l'hora d'obtenir i valorar els models, doncs ens permetria fer-nos una idea més acurada de la predicció dels nostres models i com s'apropen.

L'enunciat de la pràctica ens proposava escollir tres models lineals (o quadràtics) i dos no lineals. D'entre els models lineals treballats a l'assignatura, hem decidit escollir *Linear Regression, Ridge Regression* i *LASSO.* Dels models no lineals, hem escollit *MLP* i *Random Forest.*

Per resoldre el problema i poder fer una comparació entre els models justa, hem dividit les dades en 80% de *train* i 20% de *validation* (o *test*), ja que tenim un nombre de dades suficient per que la partició de validation sigui prou representativa amb un 20% (806 dades) i comparat els resultats obtinguts amb mètriques independents del model per tal de poder valorar correctament el rendiment de cada un.

En aquest apartat comentarem breument els models decidits a provar i alguna decisió a mencionar, així com els resultats obtinguts amb cada un d'ells. Destacar que els codis dels models es troben comentats per tal de facilitar la comprensió.

3.1. Linear Regression

Per resoldre el problema mitjançant *Linear Regression*, hem instanciat el model mitjançant la formula que relaciona el nombre de rings com una combinació lineal de les variables d'entrada, *sex* (codificat per *one hot encoding,* el que són 3 variables), *length, diameter, height, whole_weight, shucked_weight, viscera_weight* i *shell_weight.*

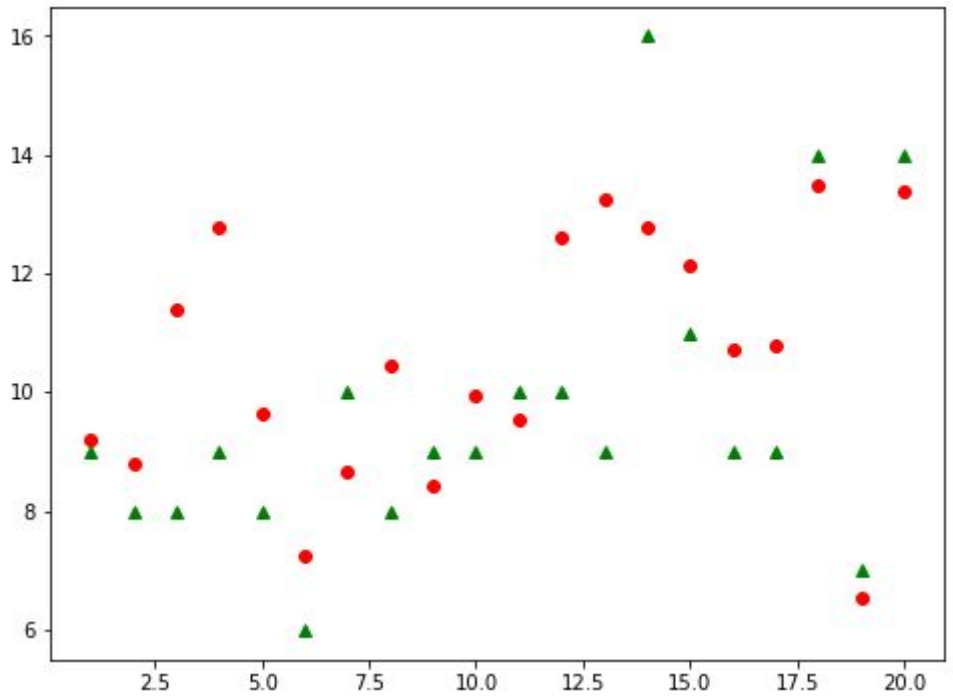

Figura 8. Valor predit (vermell) vs valor real (verd)

La figura 8 ens presenta una petita part dels resultats de validació, on es poden observar que hi ha resultats raonablement propers, pero d'altres que s'allunyen bastant del valor real. Els valors obtinguts han sigut de 1.53 de *Mean Absolute Error*, 4.06 de *Mean Squared Error* i 48.36% de R-Squared, tot aplicat sobre la part de validació.

3.2. Ridge Regression

Hem decidit usar Ridge Regression per fer una comparació del model obtingut amb el de regressió lineal sense regularitzar. Esperem, però, una precisió semblant en comparació amb el model sense regularitzar.

Per la creació del model de Ridge Regression hem necessitat fer una *Cross Validation* que ens ajudés a trobar la millor lambda. Un cop trobada, hem tornat a aplicar CV per refinar-la més i, finalment, hem obtingut el model.

Els resultats obtinguts de *MAE, MSE* i *R-Squared* han sigut de 1.52, 4.05 i 48.6% respectivament, sobre validació. Hem obtingut un *R2_LOOCV* del 50%.

3.3. LASSO

Pel model de LASSO, hem aplicat CV un cop per cercar la millor alpha. Un cop trobada, hem fet el *fit* del model un nombre alt d'iteracions, doncs l'execució és molt ràpida i ens assegurarà trobar el millor model, i hem fet les prediccions sobre el model de validation.

Intercept	4.785390695119819
male	0.08045379663168557
female	-0.0
infant	-0.7778287951284197
length	-0.0
diameter	4.229477451301426
height	5.371569648485323
whole_weight	14.349452305675698
shucked_weight	-17.167714583998155
viscera_weight	-3.085281147681697
shell_weight	4.818780174900979

Figura 9. Coeficients de LASSO

Com podem veure, amb el model de LASSO els coeficients de male i female són molt baixos, fet que ja esperàvem, doncs al fer l'anàlisis de les dades hem trobat que té poca correlació amb el resultat de rings que tingui el cargol. També es pot veure a 0 el de length.
Els resultats obtinguts de *MAE, MSE* i *R-Squared* han sigut de 1.51, 4.02 i 48.9% respectivament, sobre validació. El *R2_LOOCV* ens ha donat d'un 52%, en aquest cas.

3.4. MLP

La definició del model *Multi-Layer Perceptron* la hem dividit en dues parts:
Primer, hem fet una cerca del millor model d'una sola capa. Per fer-ho, hem considerat uns quants tamanys i els hem entrenat amb l'algorisme *lbfgs,* doncs *sgd* i *adam* no ens estaven oferint tan bons resultats, encara que aquests siguin menys costosos. La nostra primera idea era utilitzar tant la loss com el *MSE* de validació per escollir el millor valor, però hem optat finalment per usar dades de validació a l'hora de decidir quin model està aprenent millor del problema. Així doncs, finalment, només hem utilitzat el *MSE* en validació per trobar el millor model. Experimentalment, hem trobat que per 128 neurones hem trobat el millor resultat. Els R-Squared de train i validació han sigut del 59.2% i el 54.7%, respectivament. Els resultats obtinguts de *MAE* i *MSE* sobre validació han sigut de 1.398 i 3.56.

Per una altra banda, hem volgut provar a definir una xarxa amb més d'una capa per veure si aconseguíem millors resultats. En aquest cas, no hem fet cap cerca en busca d'una combinació millor. A més a més, hem volgut dividir l'entrenament en dues fases, baixant la *learning rate* a la segona. No obstant, no hem vist millores significatives, obtenint un *MAE* de 1.39, un *MSE* de 3.52 i un *R-Squared* de 60.4% en train i 55.1% en validació.

3.5. Random Forest

Per definir correctament el model de *random forest*, hem de decidir el nombre d'arbres que volem que tingui. Per fer-ho, hem anat provant de forma iterativa entre diversos valors de *trees* de forma incremental i trobant el seu oob score i el seu mse, en validation, per veure la millora que aporta augmentar el nombre d'*estimators* a la resolució del problema. Una vegada trobat, ens hem quedat amb la mitjana dels dos valors per donar-li el mateix pes a ambdós criteris de selecció, ja que, tot i que el resultat esperat seria que fossin el mateix nombre de arbres segons els dos criteris, no a totes les execucions ha estat així.

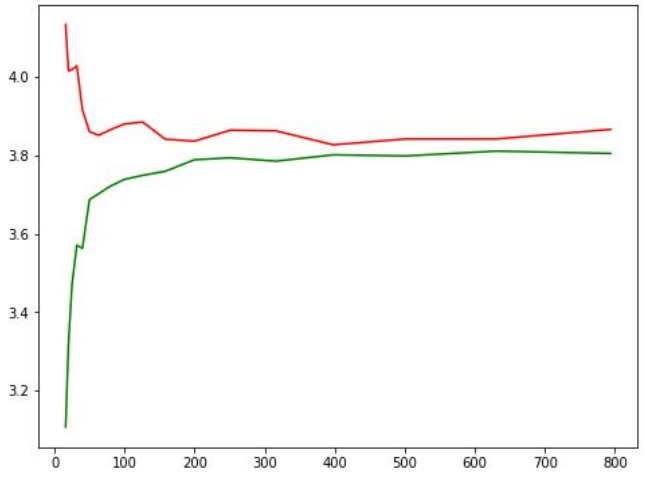

Figura 10. Oob score (verd) i MSE (vermell) en funció del nombre d'*estimators*

Com es pot veure a la figura 10, la tendència és similar amb els dos criteris. Notar que hem reescalat el valor de oob score per tal que la gràfica quedi millor ajustada.

Dels resultats obtinguts amb *Random Forest*, és important destacar el valor *R-Squared* que hem obtingut sobre el dataset de train, que ha sigut del 93%, mentre que sobre el de validació es queda al 51%. D'altra banda, hem obtingut un *MAE* de 1.42 i un *MSE* de 3.84, sobre les dades de validació.

4. Comparació entre els models

Per avaluar els nostres models, necessitem una mètrica comuna en tots ells que ens serveixi de comparatiu.

A tots els models, hem obtingut el *Mean Squared Error*, doncs tots els models usen les dades amb les mateixes magnituds i aquest ens serveix ja de comparatiu.

Addicionalment, hem volgut obtenir el *Mean Absolute Error,* doncs considerem que ens ofereix un indicatiu molt bo de la proximitat del model a la solució que hauria de donar i, de fet, ho podríem entendre com “Per quants anys s’equivoca, en mitjana”.

D’altra banda, hem vist que, generalment, per comparar resultats sobre aquest dataset, s’ha usat el *Root Mean Squared Error*, llavors l’hem obtingut tot i que creiem que no aporta massa per aquesta pràctica.

També hem volgut obtenir el *R2-Squared* per poder tenir una representació de la quantitat de variança que és capaç d’explicar el nostre model.

Addicionalment, hem calculat el *R2_LOOCV* als model lineals per poder fer una bona comparativa entre ells.

Podem dir, doncs, que tant el *Mean Squared Error* com el *Mean Absolute Error* ens serviran com a bons comparadors, encara que qualsevol de les mètriques mencionades es podrien usar. No valdria usar, però, la loss -o el score- dels models no lineals per comparar-los.

Dels models lineals, hem vist com entre ells no hi ha una diferència de rendiment a destacar. És important, però, tenir en compte que *ridge regression* ens dóna més estabilitat que *linear regression*, i que amb *LASSO* hem localitzat fins a 3 paràmetres, longitud i si és mascle o femella, sense pèrdua de rendiment, fet que realment ja s’informava al dataset.

D’altra banda, amb els models no lineals hem millorat lleugerament el rendiment, tot i que també hem augmentat de forma considerable la complexitat dels models. Entre els models d’*MLP* d’una capa o de més d’una, considerem que la elecció d’afegir capes és més eficient,

doncs la convergència ha sigut, generalment, més ràpida i ofereix els mateixos o millors resultats. D'altra banda, tot i que el model de *MLP* ens ofereix més informació sobre la variança de les dades, el model de *RF* té un temps d'entrenament notablement inferior, així que, fent un *trade-off* entre el temps i els resultats obtinguts, considerem que *RF* és la millor opció dels models no lineals.

Per decidir quin model és més adient en global, creiem que depèn dels interessos que es tinguin a l'hora d'utilitzar aquests models. Si necessitem la major precisió possible, escolliríem el model d'*MLP.* Si el que prima és un temps ràpid d'inferència, creiem que tant *LASSO* com *Ridge Regression* són igual d'adients.

5. Conclusions

Després d'haver analitzat el problema i haver provat diferents mètodes de resolució per a resoldre'l podem extreure les següents conclusions:

Dades

Pel que fa a les dades observem principalment tres característiques importants:

1. La diferència de sexe entre mascle i femella no és gairebé rellevant per a la dada objectiu, doncs el comportament de la resta de variables es veu força independent a aquesta.
2. Els infants són més difícils de predir ja que la seva correlació amb la resta de variables és molt menor que la dels altres dos sexes.
3. Degut a que arribats a certa edat els *abalone* deixen de desenvolupar-se, predir l'edat a partir dels 15 anys es fa molt complexe, el que ha estat confirmat per el descart de la variable longitud en el model LASSO.

Models

En quant als models podem concluir que els models amb major complexitat obtenen millors resultats, però amb un temps d'entrenament i inferència notablement superiors, sobretot d'entrenament.

Resultats

Finalment per el que fa a resultats, els nostres models no han sigut capaços d'obtenir un *R-Squared* per sobre del 70%, el que encaixa amb la correlació de les variables que hem vist anteriorment, on cap variable aconsegueix una correlació amb la variable objectiu major al 65%.

Podem veure, això sí, a partir del *MAE,* que el nostre millor model prediu en mitjana l'edat amb un error de 1.4 anelles, el que, tenint en compte les dificultats que trobem en les dades, que fan que tant en infants com en adults els nostres models no siguin gaire acurats, ens fa pensar que en *abalones* d'edat mitjana el nostre encert és satisfactori.

Així doncs, podem concloure finalment que les dades oferides pel dataset no ofereixen la suficient correlació com per a obtenir un gran encert a les prediccions del nombre d'anells/anys, tot i que segurament extraient les dades dels infants i dels adults amb més de 15 anys podríem obtenir un bon predictor de l'edat en una franja d'edat intermitja.

6. Bibliografia

[1] "The Abalone Dataset," *Center for Machine Learning and Intelligent Systems*. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Abalone [Accessed: 1-Nov-2020].

[2] "Ways to Detect and Remove the Outliers," *Medium*. [Online]. Available: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba [Accessed: 16-Des-2019].

[3] "All about Categorical Variable Encoding," *Medium*. [Online]. Available: https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02 [Accessed: 16-Des-2019].

[4] "Simple Methods to deal with Categorical Variables in Predictive Modeling," *Analytics Vidhya*. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/ [Accessed: 16-Des-2019].

[5] "SKLearn Documentation," *SciKit Learn*. [Online]. Available: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.neural_network [Accessed: 10-Gen-2020].

[6] "Abalone Dataset," *Kaggle*. [Online]. Available: https://www.kaggle.com/rodolfomendes/abalone-dataset/kernels [Accessed: 29-Des-2019].

[7] Pràctiques del Laboratori d'Aprenentatge Automàtic. FIB., 2019-2020.[Accessed: 1-Nov-2020].