Inés Broto Clemente
Arnau Turch Ferreres
Víctor Novelle Moriano

Classificació de tumors a les glàndules mamàries

1. Descripció del projecte i presentació de les dades

Des d'un inici, el principal objectiu del nostre grup consisteix a treballar un tema relacionat amb l'àmbit mèdic com ja vam comentar a l'hora d'escollir les dades de treball i, entorn aquest tema, estavem especialment interessats en classificar els fetus en tres estats segons com s'està duent a terme la seva gestació i la classificació de tumors de cancer de pit. Ambdues opcions ens semblaven molt interessants i finalment ens vam decantar pel treball sobre els fetus ja que ens semblava d'una alta utilitat pràctica, ja que determinar de manera avançada anomalies en la gestació pot ajudar a prevenir futures malalties i complicacions.
Abans de començar a treballar directament amb les dades, com era d'esperar, vam haver de fer una mica de recerca sobre el camp de l'anàlisis. Malgrat la base de dades incloïa una breu descripció de totes les variables, vam haver de posar cara i ulls a tot aquell grapat de sigles per tal de poder començar a fer el treball de pre-processament. Aquest primer contacte amb els tecnicismes del camp va resultar interessant però a l'hora limitat tenint en compte els problemes als que vam haver d'enfrontar-nos quan vam començar a treballar amb les dades.

En primer lloc, al descarregar-nos les dades de la pàgina esmentada a les referències i realitzar una visualització prèvia,ens hem trobat amb el problema de la interpretació de les variables. Problema que vam poder solucionar després de veure que el nostre *dataset* incloïa algunes variables exclusivament identificadores que no constaven a la documentació.

De totes maneres, aquesta no va ser la principal raó del canvi de rumb, quan vam començar a treballar amb les dades ens vam adonar que teníem una enorme quantitat d'observacions atípiques que fàcilment podien esbiaixar els nostres resultats. En un principi vam considerar treure-les però això limitava molt el nombre de dades així com la representació equitativa dels grups.

Per aquesta raó vam replantejar-nos el projecte. Empesos per un estudi que evidenciava la gran quantitat de valors atípics presents a la nostra base de dades (veure referències[5]), vam decidir canviar de matèria per dedicar-nos a l'altre opció que des d'un principi ens havia cridat l'atenció.

D'aquesta manera podríem treure molt més suc al seu anàlisi i, lluny de tenir limitacions en aspectes que queden fora dels objectius de l'assignatura, podríem centrar-nos en aprofundir en tots els tipus de classificadors poposats a classe.

La nostra nova base de dades treballa amb 9 variables ordinals (codificades com a variables enteres o contínues en funció del mètode de treball) per mitjà de les quals s'intentarà predir si un tumor a les glàndules mamàries és benigne o maligne. Les dades provenen d'anàlisis clínics periòdics a 699 pacients sobre els quals s'estudiaven certes característiques dels tumors i eren avaluades de l'1 al 10 pels doctors. Aquestes variables són les següents:

1. Gruix del grup tumoral 1 - 10
2. Uniformitat de la mida de la cèl·lula 1 - 10
3. Uniformitat de la forma de la cèl·lula 1 - 10
4. Adhesió marginal 1 - 10
5. Tamany de la cèl·lula Epithelial 1 - 10
6. Nuclis nus 1 - 10
7. Cromatina suau 1 - 10
8. Nuclèol Normal 1 - 10
9. Mitosis 1 - 10
10. Classe benign, malignant

Tal i com havíem fet amb la base de dades anterior, abans de començar a treballar amb la base de dades crua, hem realitzat un petit treball de recerca sobre el camp per entendre el significat i la influència que poden tenir totes les nostres variables als resultats que obtenim.

3. Exploració i pre-processament de les dades

Al donar un primer cop d'ull a la base de dades hem vist que podem desfer-nos de la primera columna tenint en compte que simplement identifica cada pacient. Anant una mica més enllà, hem comprovat que tenim 16 dades faltants codificades com a "*?*" però totes pertanyen a la mateixa variable *(BN)*. Això últim ens ha fet decidir que la imputació és una bona solució al problema ja que l'eliminació de totes les observacions no és necessària podent predir fàcilment el valor de les dades mancants per mitjà del mètode del veïnatge proper. A més a més d'imputar les dades, hem hagut de recodificar els valors de la variable *BN* per tal d'evitar futurs problemes ja que la base de dades encara considerava la possibilitat que aquesta variable prengués el valor "*?*" i això podia generar un conflicte a l'hora de modelar en funció d'aquesta variable, així com els valors de la variable classe, que els hem recodificat com a "*benign*" i "*malignant*" per tal de simplificar la interpretació dels resultats. A més a més, hem comprovat que tenim gairebé el doble d'observacions de tipus benigne que maligne. Això en principi no hauria de resultar un

problema però no està de més tenir-ho present per si en algun moment observem prediccions estranyes o clarament esbiaixades ja que al treballar amb grups de dades escollits aleatòriament podríem tenir la mala sort de sobre-representar o infra-representar una classe. Per evitar aquest tipus de situacions realitzarem més d'una modelització del mateix tipus de manera iterativa per tal de reduir al mínim la probabilitat d'obtenir particions extremes.

Arribats a aquest punt, ja tenim la base de dades llesta per començar a treballar, i per això el primer que volem fer és visualitzar en la mesura del possible la informació que tenim. Disposem de diverses funcions que ens poden ajudar a fer aquesta exploració una mica més profunda a les nostres dades però no totes casen amb el nostre estudi. La funció pairs, que en algun altre cas ens podria ser útil per analitzar la relació entre variables dos a dos, no ens aporta informació en aquest cas perquè totes les nostres variables son ordinals. Realitzar un anàlisis de correspondències múltiple tampoc ens convé ja que al treballar amb variables ordinals no té sentit estudiar-ne les correlacions. Així doncs, procedim a realitzar un anàlisis de components principals considerant que les nostres variables prenen valors enters d'entre els reals.

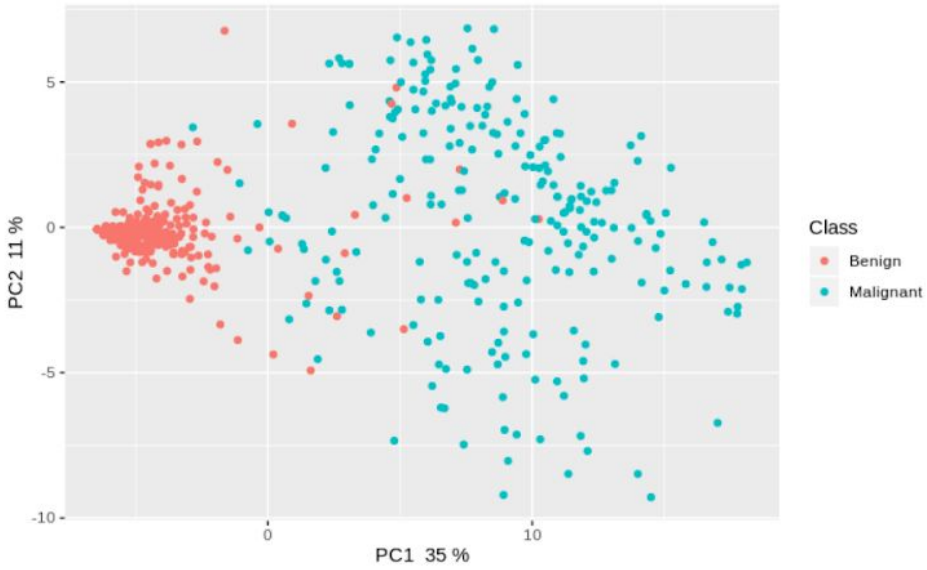

Figura 1: Biplot de les components principals

Tal i com podem veure, amb les dues primeres components principals de les dades ja queden força separades tot i només estar explicant un 46% de la variància. Això ens indica que les nostres dades estan clarament diferenciades i que per tant obtindrem, segurament, resultats molt satisfactoris.

Una altra manera de visualitzar les nostres dades i assegurar-nos que pertanyen a dos grups diferenciats és realitzar un anàlisis cluster. Com que els resultats d'aquesta metodologia van

estretament lligats a la manera de realitzar aquest anàlisis hem optat per realitzar-lo prenent com a mesura de distància la distància euclidiana i treballant amb diferents linkatges.

En el cas del linkatge complet o per veïnatge llunyà, la separació en dos grups no és gaire bona ja que hi ha 102 observacions que queden mal classificades si tallem el dendrograma pel nivell 2 i prenem aquesta classificació, cosa que correspon a un 85,5% d'encert. Això, per tant, no ens dona una clara confirmació que les dades estiguin clarament bipartides. Per contra, al treballar amb el linkatge Ward, els resultats són molt més esperançadors. Ambdós grups queden clarament diferenciats amb un encert del 96,4% prenent la classificació en clusters de nivell 2. Amb tot això podem concloure que la nostre base de dades conté clarament informació de dos grups diferenciats i per tant és convenient proseguir amb el desenvolupament de models per tal de predir la condició de tumors mamaris a partir de les variables ja esmentades.

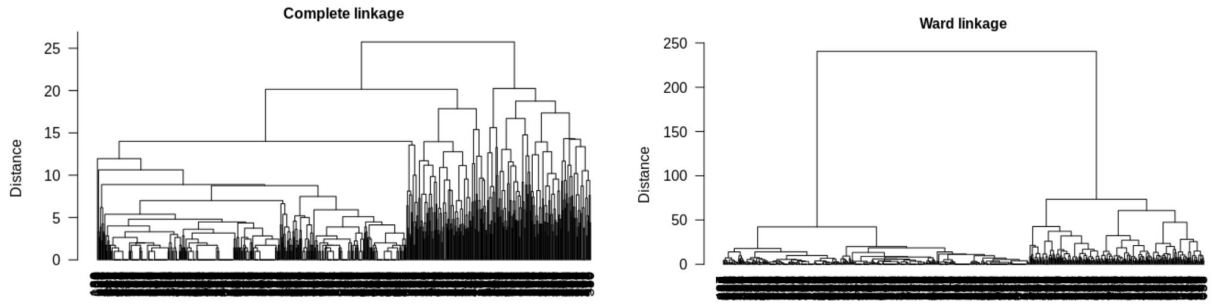

Figura 1. Plot dels dos mètodes clusters utilitzats.

A continuació, seguim amb la visualització de les dades de la mà dels histogrames per observar amb quina freqüència obtenim les respostes de cadascuna de les variables i quina relació podem preveure amb la classificació de tumors en benigne i maligne. No hem d'oblidar que a la nostra base de dades tenim un 66% de representació de casos benignes i un 34% de representació de casos malignes, per tant les dues classes no estan igualment representades.

Observant els histogrames podem afirmar que en general les respostes amb índexs baixos corresponen a casos classificats com a benignes i viceversa. Cal destacar que no tenim mostres amb valor 9 per la variable *SECS,* cosa que probablement haurem de tenir en compte a l'hora de desenvolupar els nostres models.

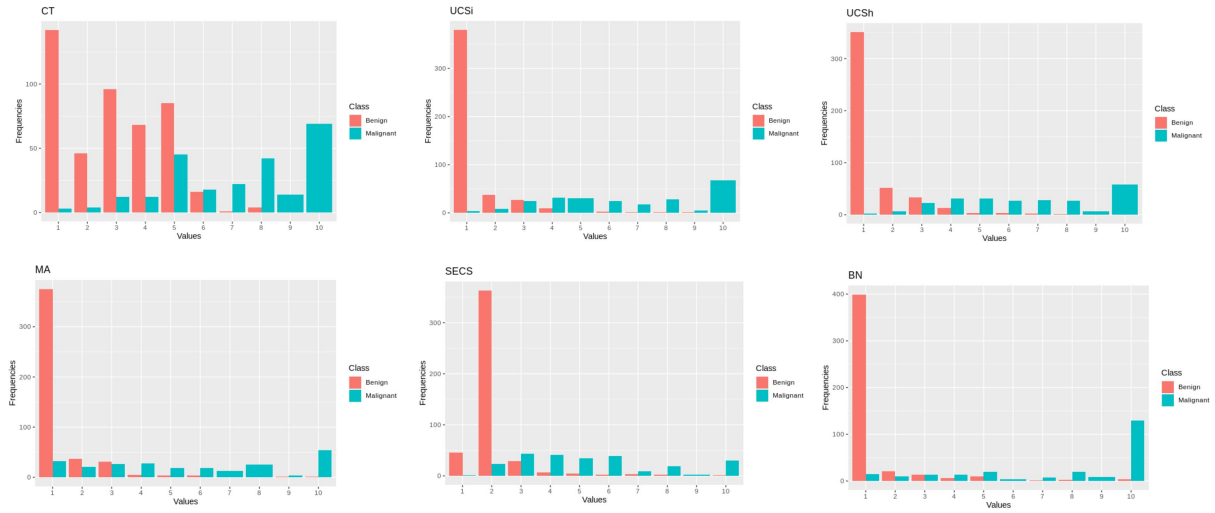

Figura 2. Histogrames de les 6 primeres variables (Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size i Bare Nuclei)

El fet que les distribucions de casos benignes i malignes siguin notablement diferents per totes les variables ens fa pensar que les nostres dades seran fàcilment classificables i serem capaços de construir models amb molt bones capacitats de predicció.

4. Re-mostreig i modelatge

Un cop realitzada la part de pre-processament de les dades, vam decidir quin procediment de *resampling* escolliriem per tal d'obtenir el millor model. Degut a que la selecció del millor model, així com l'estimació mostral del *true error* es realitzen de manera simultània, vam escollir una *three-way* partition. En aquesta, vam reservar un **33 %** de les dades per a test i la resta per a entrenament i validació.

Respecte a la divisió entre validació i entrenament es va optar per implementar un *ten times ten-cross validation.* És a dir, realitzem deu vegades un *ten-fold cross validation* modificant la partició inicial en cadascuna d'elles. L'elecció d'aquest mètode de *resampling* ve determinada per dos grans factors.

El primer és que, degut a ser un mètode de mostreig de les dades més complex que *random subsampling* o el *holdout method*, ens permet obtener una estimació més precisa dels errors de validació, realitzant la mitjana dels trobats en cadascuna de les particions aplicades.

En primer lloc vam plantejar intentar realitzar un model lineal generalitzat, més conegut com glm, i és aquí on ens vam trobar un dels principals problemes que hem enfrontat: el fet que les

dades siguin ordinals. Això en realitat va acabar sent un avantatge, ja que ens va permetre treballar des del punt de vista categòric però també continu, ampliant així el ventall de possibles metodologies de treball.

De totes maneres, en un principi vam considerar les variables com a categòriques i aquest fet tan subtil ens va limitar molt a l'hora de classificar, ja que si un nivell de la variable no és gaire freqüent, és molt possible que no estigui present en les tres divisions, i en el cas de que el model no hagi estat entrenat amb aquesta variable, a l'hora d'intentar predir amb les dades de test o validació utilitzant una mostra on aquest nivell del factor que no està present en el training set, el model no és capaç de classificar l'individu.

Després de veure que treballar només des del punt de vista categòric no era una opció viable vam reconsiderar la naturalesa de les nostres variables per adonar-nos que realment eren ordinals i de valor enter i per tant les podíem tractar com si fossin nombres reals. Malgrat a priori això podria semblar una barbaritat, en el nostre cas no ho és ja que per cada variable tenim 10 possibles nivells ordenats. Els resultats obtinguts són esperançadors ja que, malgrat estiguem fent una predicció de valors reals, aquests no són més que les 10 categories presents a cadascuna de les nostres variables, llavors una predicció numèrica corresponent a un valor és perfectament vàlida com a una predicció.

4.1. GLM

Un cop superat aquest entrebanc vam realitzar un total de tres estudis amb glm, utilitzant els tres tipus de links més famosos i important dins de la familia binomial: el logit, el probit i el clog-log. L'elecció d'aquesta família es deu a que al cap i a la fi, el que nosaltres estem intentant fer és classificar totes les observacions en dos grups, cosa que correspon a comptar, dins un total d'observacions, quantes són de tipus 1 (benigne, en aquest cas) i quantes de tipus 2 (maligne). Així doncs, la distribució de probabilitat que més escau per dur a terme aquest anàlisis és la binomial.

Tots tres models ens van brindar uns errors de validació bastant raonables i satisfactoris, sobretot el link probit que és el que va assolir un percentatge d'error menor, per tant, serà un dels principals candidats a ser el nostre model.

4.2. Naïve-Bayes

En segon lloc vam utilitzar el mètode Naïve-Bayes, un mètode molt fàcil d'implementar gràcies en primer lloc a la llibreria "NaiveBayes" i en segon lloc a que el fet de que les dades siguin factors no suposa cap tipus d'inconvenient ja que cal recordar que el Naïve Bayes continuarà

funcionant al predir o classificar una mostra que presenti una variable desconeguda pel model si s'utilitza la correcció Laplaciana, utilitat que la funció per generar el classificador Naïve-Bayesià disposa. Punt a favor d'aquest mètode.

Sense deixar de utilitzar el 10 times 10-fold cross validation vam calcular l'error de validació i ens va donar un valor envejable per qualsevol classificador que es vulgui fer respectar, i clarament, al costat del glm probit, estarà a la llista per triar el nostre model final.

4.3.KNN

En tercer lloc ens vam centrar a realitzar el mètode dels k-nearest-neighbours, ja que durant el pre-processament de dades, al realitzar diferents tipus de clusters, els resultats van ser bastant satisfactoris. Es podien veure clarament dos grups ben diferenciats i aquest fet ens va portar a pensar que potser utilitzar les distàncies entre les mostres ens proporcionaria una informació bastant útil i significativa.

Utilitzem la llibreria de R "knn" per tal de calcular l'error de validació d'aquest tipus de model i valorar el seu rendiment. Per tal de reforçar el valor d'aquest error, es raonable realitzar un knn Leave One Out Cross Validation, més conegut com LOOCV.

Pero fins aquest punt tots els algoritmes de knn els haviem realitzat utilitzant un nivell de veïns arbitrari, que havíem preestablert nosaltres sense cap tipus d'evidència de si era el nivell òptim. Per trobar el nombre de veïns més adient es pot realitzar aquest knn LOOCV de manera iterativa, iterant sobre el nombre de k per tal d'assolir el nostre propòsit esmentat anteriorment.

Vam realitzar aquesta recerca i la conclusió que vam extreure que el nombre k òptim es 6. En un principi el fet que sigui parell no ens acaba de convencer ja que això obre la porta a possibles empats entre veïns benignes i malignes però tenint en compte que l'error utilitzant 6 veïns és considerablement més baix ens sembla raonable escollir aquest valor i resoldre els empats de manera aleatòria. Segurament el nombre d'empats serà baix ja que les dades estan força ben diferenciades i per tant ens compensa prendre 6 com el valor òptim del paràmetre. Si més no, si no volem resoldre empats de manera aleatoria sempre podriem seleccionar 5 veïns, que és el nombre imparell de veïns amb mínim error, a costa d'augmentar lleugerament l'error de validació, passant de 0.02564103 a 0.02991453.

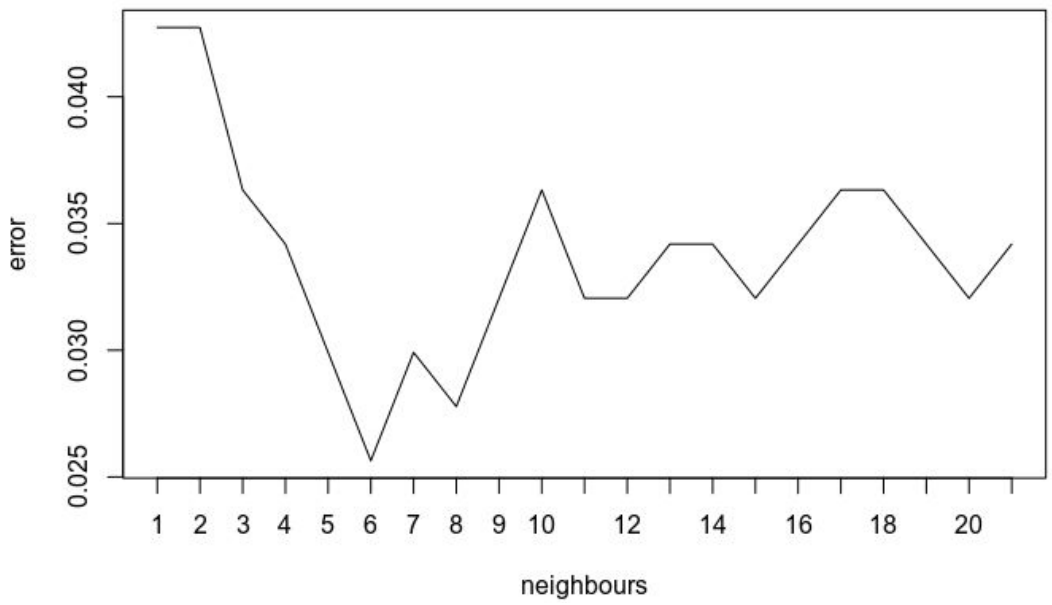

Figura 3. Error del classificador segons el nombre de veïns utilitzat

4.4. QDA i LDA

Finalment, ens vam disposar a realitzar els classificadors gaussians generatius, tant QDA com LDA.

Primer de tot generem el nostre classificador LDA, pero abans de tot realitzem un plot d'aquest per observar que tal i com podiem i voliem esperar, es veuen els dos grups clarament diferenciats, aquest classificador funcionarà molt bé. Un cop feta aquesta petita observació sol va faltar mesurar la seva efectivitat i rendiment en la classificació.

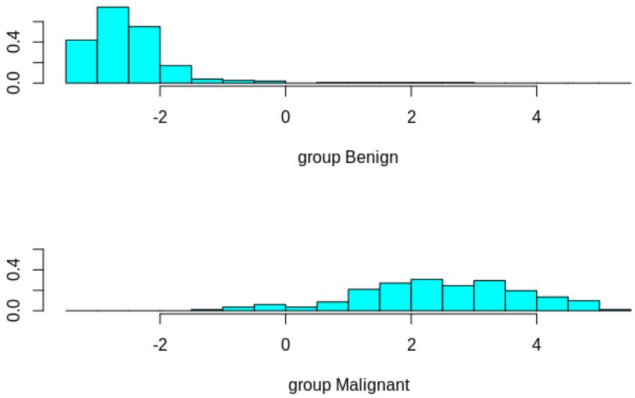

Figura 4. LDA

Seguidament realitzem el QDA, i de la mateixa forma calculem el seu error de validació.

4.5. Artificial Neural Network

Crear una xarxa neuronal amb R no és molt complex ja que disposem de la funció "nnet", el que sí és més complex és decidir el número de neurones o valor de regularització utilizar. Per això en primer lloc fixarem un valor de regularització 0 i observarem per quin valor de neurones obtenim el mínim error de validació utilitzant un 10x10 CV. Per altra banda realitzarem el mateix procediment pero en aquest cas fixarem el número de neurones en un valor alt, i aplicarem diferents valors de regularització utilitzant el mateix mètode de validació.

En el primer cas utilitzant 2 neurones i un decay 0 obtenim un 0.03351045 d'error. En el següent gràfic podem comprovar l'evolució de l'error de validació utilitzant aquest mètode.

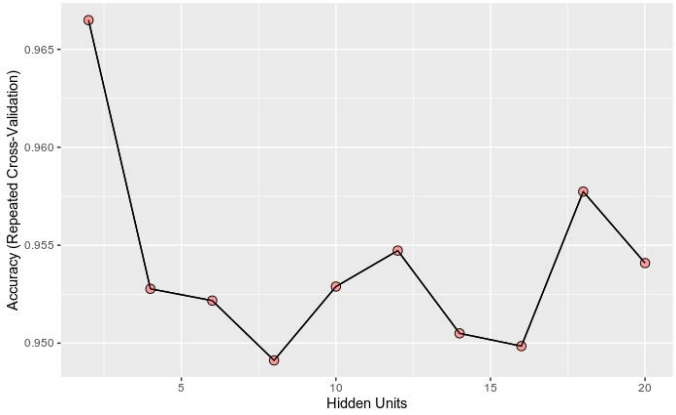

Figura 5. Error vs número de neurones ocultes

En el segon cas, fixant el valor a 20 neurones obtenim 0.02884752 d'error amb un valor de decay 1.

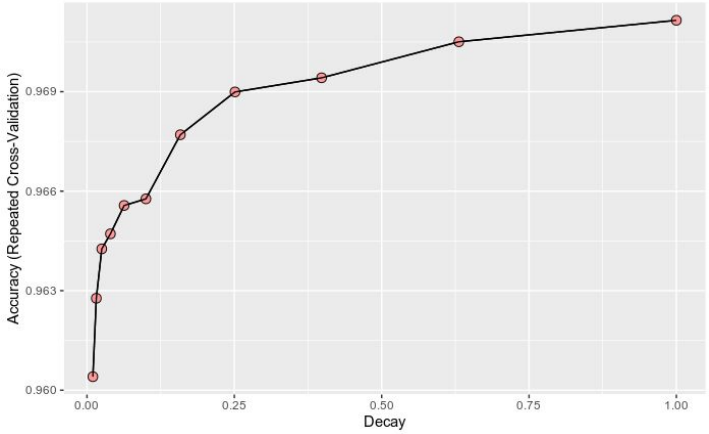

Figura 6. Error vs decay

Amb això ens quedarem amb la xarxa neuronal del segon cas, 20 neurones i un valor de decay igual a 1 ja que és qui mostra un error de validació més baix.

4.6. Random forests

Per últim, també hem implementat un mètode ensamblador, un dels més populars, com és el *random forest.* Tal i com ja sabem gràcies a les classes de teoria, aquests mètodes consisteixen en entrenar individual learners (millors que l'atzar), i posteriorment combinar les seves prediccions.

Per a la implementació d'aquest mètode, utilitzarem la funció *randomForest,* proporcionada a les sessions de laboratori.

El primer pas en l'aplicació d'aquest mètode consisteix en seleccionar quina sera la quantitat d'arbres a utilitzar. Per tal de realitzar aquest procediment, provarem diverses mides d'arbres i mirarem el seu respectiu *Out Of Bag Error.* Degut a que utilitzem l'*OOB* queda implicit que un mètode de *Bagging*, ha estat utilitzat amb l'objectiu d'aconseguir disminuir la variància en els arbres de decisió, que són els individual learners en el nostre mètode ensamblador i, com sabem, són mètodes inestables, amb l'objectiu de millorar la capacitat de predicció.

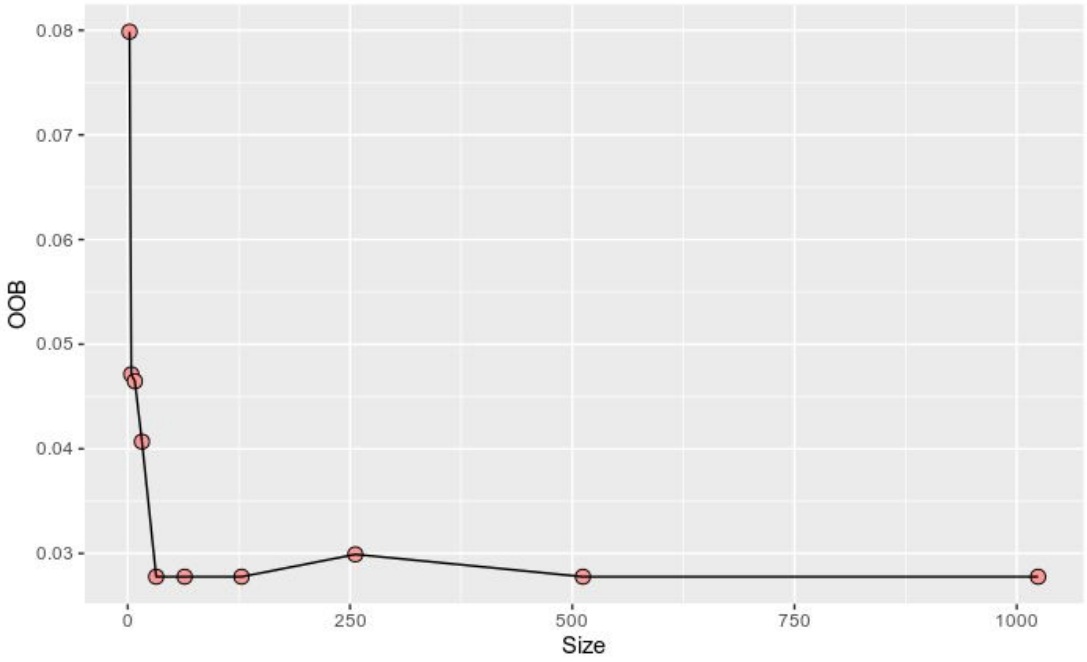

Figura 7. OOB vs size

Com podem comprovar, en el nostre cas el nombre d'arbres que ens proporciona un menor *OOB* error és 32 , sent per tant l'escollit per, posteriorment, aplicar una 10x10CV i comparar el resultat amb la resta de models presentats en la pràctica.

5. Resultats preliminars i comparació

En la següent taula es mostren els mètodes utilitzats juntament amb el seu error de validació utilitzant el ten times ten cross validation:

MÈTODE	ERROR DE VALIDACIÓ
GLM Binomial (link logit)	0.03419981
GLM Binomial (link probit)	0.03270583
GLM Binomial (link cloglog)	0.04082331
Naïve Bayes	0.02565217
Knn 10 neighbours	0.03266883
Knn LOOCV, 10 neighbours	0.03205128
Knn, 5 neighbours (odd classification)	0.02991453
LDA	0.03267808
QDA	0.04918131
Random Forest	0.02864015
Neuronal Network	0.02884752

Com observem el millor mètode per classificar és el Naïve Bayes ja que és el que ens ha donat un error de validació menor seguit dels mètodes Random Forest i Neuronal Network,mètodes notablement més complexes en comparació amb la resta.

De totes maneres, els resultats són molt satisfactoris ja que en cap cas es supera el 5% d'error, tal i com havíem predit amb l'exploració de les dades, on havíem pogut veure que les dues classes quedaven clarament diferenciades en funció dels valors que prenien les variables.

Respecte al knn, podem observar que l'error de predicció obtingut s'aproxima als models anteriorment mencionats, sent lleugerament superiors. Tal i com s'ha indicat anteriorment, s'ha optat per escollir el model amb veïns imparells per a evitar empats. Com és lògic, l'elecció d'un nombre de veïns més pròxim a l'òptim per a les dades del nostre problema, dona millors resultats que si escollim un nombre genèric (K = 10). A més a més, en aquest procediment, hem pogut comprovar com al realitzar particions i realitzar prediccions a partir d'aquelles particions de manera iterativa per obtenir promitjats dels errors ens dóna errors diferents a fer aquest procediment només una vegada (partició train-validation única) o a fer el cas extrem d'aquesta metodologia, el LOOCV.

Pel que fa als models lineals generalitzats, el que ens dóna un millor error de validació és el de link probit, i això és segurament degut a que utilitza una distribució normal acumulativa que té menys en compte els valors extrems.

Per últim, pel que fa als mètodes discriminants lineals i quadràtics, obtenim millors classificacions amb el lineal, cosa força positiva ja que suposa menys complexitat.

6. Model final i estimació de l'error de test

Per últim, i després d'acceptar el mètode de Naïve Bayes com el més adient per tractar amb el nostre problema de classificació, només ens falta fer una predicció de l'error de test amb les dades que havíem reservat des d'un principi. Els resultats fins ara han estat molt convenients, doncs hem obtingut valors de l'error de validació considerablement petits (un 2,6% en el cas de Naïve Bayes), però el fet d'estar calculant aquest error amb les mateixes dades amb les quals s'ha entrenat el model fa que puguem estar sobrestimant-lo.
Ja des d'un inici havíem separat la nostra base de dades en ⅔ de dades d'entrenament i ⅓ de dades de test. Aquestes últimes entraran en joc en aquest últim apartat on, amb la metodologia Naïve Bayes, intentarem predir la naturalesa de cada observació de test a partir dels valors de cada variable i compararem els resultats amb la naturalesa real. D'aquesta manera, veurem com actua el nostre model davant d'unes dades noves per ell, la situació en la que ens trobaríem si realment volguèssim classificar tumors a partir de les mesures d'un professional.

Els resultats obtinguts són molt satisfactoris, l'error de test és tan sols d'un **3,03%** i podem afirmar que: en primer lloc, aquest mètode és molt adient per tractar aquest tipus de problema i, en segon lloc, les dues classes del nostre problema són clarament diferenciables a partir de les variables que tenim.

7. Conclusions

Tal i com hem pogut comprovar al llarg de tota la pràctica, una de les coses a destacar és que la nostra base dades és molt adient per treballar des del punt de vista de la classificació amb tota la metodologia que hem estudiat durant el curs. La naturalesa benigna o maligna dels tumors a les glàndules mamàries queda molt ben explicada per les diferents característiques representades en forma de variables i això es tradueix en un molt bon funcionament de tots els mètodes implementats. A més a més, cal recalcar que en especial la metodologia Naïve Bayes ens dona uns resultats molt bons tenint en compte la poca complexitat que implica. Els mètodes de Random Forest i Neuronal Network també ofereixen bones qualitats però són considerablement més complexes, sobretot a nivell computacional.
Finalment els resultats són d'allò més satisfactoris, doncs hem trobat un model senzill capaç de classificar noves dades amb un percentatge d'error molt baix. Arrel d'això ens ha passat pel cap que per anar un pas més enllà podríem treballar buscant minimitzar la probabilitat de cometre errors a l'hora de predir casos positius (reduir al màxim els falsos negatius) encara que això signifiqués acceptar tenir algun fals positiu més. Tenint en compte el camp en el que estem treballant (oncologia) això resulta molt adient.

Pel que fa a la vessant personal, tenim la sensació que aquesta pràctica ens ha servit per assentar els coneixements que durant el curs havíem tractat des d'un punt de vista teòric. El fet de veure en acció totes aquestes metodologies i haver d'afrontar i solucionar els petits entrebancs que van sorgint al treballar amb cada mètode ens ha fet entendre molt millor el seu comportament i la seva utilitat, doncs hem estat capaços de desenvolupar diferents classificadors de tumors benignes o malignes a partir d'un conjunt reduït de dades i amb molt bons resultats. A més a més, tot el treball des del punt de vista de programació ha estat un pas necessari per, més enllà de relacionar la part teòrica i pràctica de l'assignatura, tenir eines i agafar experiència de cara a poder realitzar anàlisis d'aquest estil en un futur.

Hem obtingut la nostra base de dades a partir d'un dels repositoris proporcionats pels professors, en concret el *Machine Learning Repository UCI* a través de la pàgina *OpenML* (http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29). Es tracta d'un conjunt de dades obtingudes dels hospitals de la Universitat de Wisconsin, en especial pel Dr. William H. Wolberg, qui de manera periòdica pa prendre les dades dels casos clínics que ell mateix va portar del gener de 1989 fins al novembre de 1991.

8. Referències:

1. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

2. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

3. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

4. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

Pel que fa a les primeres dades que amb les que vam treballar, volem mencionar que van ser obtingudes de la mà dels doctors J.P. Marques de Sá (Institut d'Enginyeria Biomèdica d'Oporto, Portugal), J. Bernades (Facultat de Medicina de la Universitat d'Oporto, Portugal) i Aryes de Campos (Facultat de Medicina de la Universitat d'Oporto, Portugal) i vam accedir elles de nou per mitjà del *Machine Learning Repository UCI,* a través de la pàgina *OpenML* (https://www.openml.org/d/1560*)*. A més a més, l'estudi efectuat per Shomona i R. Geetha ens va empènyer a prendre la decisió de canviar el rumb del nostre projecte ja que en ell es feia evident el problema de l'alta presència de valors atípics que ens obligava a treballar amb un nombre de dades molt reduït.

1. Gracia Jacob, S. and Geetha Ramani, R., 2020. *Evolving Efficient Classification Rules From Cardiotocography Data Through Data Mining Methods And Techniques*. [online] ResearchGate. Available at: <https://www.researchgate.net/publication/267985491_Evolving_Efficient_Classification_Rules_from_Cardiotocography_Data_through_Data_Mining_Methods_and_Techniques> [Accessed 31 March 2020].