Model de predicció del consum de cànnabis

Alex Ferrando de las Morenas
Carlos Hurtado Comín
Maria Ribot Vilà

22 de juny 2020

Índex

1 Objectius

L'objectiu últim d'aquest projecte és familiaritzar-nos amb l'estudi d'una base de dades fent servir mètodes d'aprenentatge automàtic. Això inclou el pre-processament de les dades, així com un estudi previ per tal d'entendre les dades de què disposem i després escollir els mètodes correctes en quant a tipus i distribució de les dades. Per fer això, a les següents pàgines s'estudiarà una base de dades que conté dades demogràfiques de 1885 persones, així com els seus resultats en exàmens de personalitat i caràcter. També es disposa de dades relacionades amb el consum de diverses drogues. Encara que tenim informació d'un gran ventall de drogues, hem volgut fer un estudi del potencial d'un individu a ser consumidor de cànnabis, una droga il·legal, a partir de dades demogràfiques, de personalitat i del consum de drogues legals i socialment normalitzades com són l'alcohol i la nicotina. És per això que no es tindran en compte l'abús d'altres drogues en aquest estudi.

2 Descripció del conjunt de dades

2.1 Recopilació de les dades

La base de dades amb la que treballarem [1] ha sigut extreta del repositori de dades per aprenentatge automàtic *UC Irvine Machine Learning*. La base de dades *Drug Consumption* és original d'un estudi del departament de matemàtiques de la universitat de Cornell [2] on s'estudien cinc factors de la personalitat i com es relacionen amb el risc de consum de droga. La creació de la base de dades va córrer a càrrec de Elaine Fehreman i Vincent Egan entre el març de 2011 i març de 2012 [3]. La recol·lecta de les dades es va fer amb l'eina d'enquestes *Survey Gizmo* que permet obtenir dades amb anonimat màxim, especialment rellevant pel tema específic que s'està estudiant. Es van aconseguir 2051 respostes, 166 de les quals no van satisfer els criteris de validació i per tant es van retirar. Es va acabar acceptant les respostes d'un total de 1885 participants. Un dels criteris que es va introduir per detectar afirmacions exagerades sobre el consum de substàncies va ser la introducció d'una droga fictícia per tal d'eliminar aquells qui afirmaven haver-la consumit.

El mètode de mostreig utilitzat s'anomena *Snowball* i es tracta d'una tècnica utilitzada sobretot per fer estudis estadístics de poblacions difícils d'accedir amb altres maneres d'aconseguir mostres poblacionals. La tècnica es basa en garantir l'anonimat dels participants i que ells mateixos siguin els qui enviïn l'enquesta als seus contactes de manera encadenada. El fet que l'enquesta pugui arribar a la població desitjada pot dur a biaixos en els models realitzats; les mostres no es poden considerar independents ja que el fet de compartir l'enquesta entre grups d'amics pot produir una representació no equitativa de la mostra poblacional i afectar a les probabilitats a priori del model. Aquest fet es pot veure clarament amb les diferents variables explicatives recollides com ara països d'origen o edat on la mostra recollida es concentra fortament en determinants grups poblacionals.

2.2 Contingut de la base de dades

Un cop analitzat l'origen i característiques de la recopilació de les dades podem analitzar el propi contingut de la base de dades. Com hem comentat anteriorment

es disposa d'un total de 1885 observacions sense cap dada mancant.

1. **Edat**: Observem una població majoritàriament jove, dividida en els següents intervals: 18-24 years (643; 34.1%), 25-34 years (481; 25.5%), 35-44 years (356; 18.9%), 45-54 years (294; 15.6%), 55-64 (93; 4.9%), and over 65 (18; 1%).

2. **Gènere**: male (943; 50%), female (942; 50%).

3. **Educació**: professional certificate or diploma (271, 14.4%), undergraduate degree (481; 25,5%), master's degree (284, 15%), and doctorate (89; 4.7%), college without diploma (506; 26.8%); left school at the age of 18 or younger (257; 13.6%).

4. **País d'origen**: Majoritàriament països anglosaxons (93.7%), hi ha informació significativa dels següents: the UK (1044; 55.4%), the USA (557; 29.5%), Canada (87; 4.6%), Australia (54; 2.9%), New Zealand (5; 0.3%), Ireland (20; 1.1%), Other(118; 6.3%)

5. **Mesures de personalitat**: Es va utilitzar el qüestionari *Revised NEO Five-Factor Inventory (NEO-FFI-R)* que és una mesura significant dels dominis bàsics de la personalitat. Es pot fer el test a la pàgina web: `https://www.truity.com/test/big-five-personality-test`. El test intenta fer una mesura quantitativa aproximada de cinc factors de la personalitat:

   i. *Neurotisme (N)*: tendència a llarg termini a experimentar emocions negatives: nervis, tensió, ansietat o depressió.

   ii. *Extraversió (E)*: es manifesta en ser extravertit, proper, actiu, assertiu, alegre i la cerca de característiques estimuladores.

   iii. *Obertura a noves experiències (O)*: apreciació general per l'art, idees poc convencionals o usuals i interessos eclèctics, imaginatius i creatius.

   iv. *Agrat (A)*: dimensió de relacions interpersonals: altruisme, confiança, modèstia, amabilitat, compassió i cooperació.

   v. *Consciència (C)*: tendència a ser ordenat i algú de qui es depèn, conviccions fortes, persistent, eficient.

   També s'utilitza *Barratt Impulsiveness Scale (BIS-11)* que és un qüestionari disponible a `https://scielo.conicyt.cl/pdf/rchnp/v51n4/art03.pdf` de 30 preguntes que mesura la construcció de comportament impulsiu amb tres subescales:

   i. *Impulsivitat motora*: Propensió a actuar sense pensar.

   ii. *Impulsivitat d'atenció*: Poca concentració i intrusions al pensament.

   iii. *Poca planificació*: Falta de consideració de les conseqüències.

   Per acabar, també s'usa *Impulsiveness Sensation-Seeking (ImpSS)* com a eina de mesura del comportament, sobretot de la tendència a explorar noves sensacions. Totes les mesures de personalitat representen un conjunt de valors enters.

6. **Consum de drogues**: La base de dades conté 18 factors que descriuen el consum de diferents drogues en els perfils estudiats, degut a la quantitat de informació donada per l'enquesta i l'objectiu de la nostra pràctica hem reduït el nombre de categòriques que ho defineixen. Per poder fer un estudi focalitzat hem conservat les variables de consum de drogues legals (alcohol i nicotina, és a dir tabac, *vappers* i semblants) com a variables predictores per a la nostra variable objectiu que és el consum de cànnabis. Obtenim de cada factor 7 nivells que abasten des del 0 (no ha consumit mai) - fins el 6 (consumit avui) tot i que per la variable objectiu separem entre NON-USER (no consumit en l'últim any) i USER (consumit al menys un cop en l'últim any).

Cal destacar que aquestes categories venen codificades a la base de dades com a valors reals en canvi de com a factors o enters tot intentant aproximar la distribució normal. Sigui la categoria i-èssima d'una variable explicativa, es codifica de manera:

$$t_i = \phi^{-1}(\sum_{j=1}^{i} p_j)$$

Després de preprocessar les dades i estudiar el funcionament, entrenament i temps d'execució vam decidir fer l'estudi previ amb les dades transformades a les categories corresponents ja que de l'altre manera perdem totalment la interpretabilitat de les dades i ens trobaríem amb un anàlisi incongruent. Així doncs, amb aquesta transformació podríem obtenir variables explicatives amb sentit però només des dels factors ordinals. Aquesta transformació a les dades, per tant, no ens aporta cap avantatge significatiu respecte a treballar amb les variables com a factors. El que sí que podem intentar és, gràcies a la codificació de les variables, implementar models classificatoris per a variables continues com ara QDA i LDA tot i que esperarem errors majors.

Així doncs, per plantejar els nostres models tindrem un total de sis variables explicatives categòriques, set variables explicatives que prenen domini en els naturals i la variable resposta binaritzada que indica si la persona és consumidora de cànnabis o no.

3 Anàlisi

Abans d'entrenar cap model per a la predicció del consum de cànnabis en persones adultes hem de fer una anàlisi exploratori de la base de dades per garantir el bon funcionament dels models predictors. En primer lloc, analitzarem aquells aspectes que poden distorsionar tant l'anàlisi previ com l'exploració del model.

Abans d'implementar els mètodes de classificació estudiarem la relació entre les variables predictores amb la propensió de ser o no consumidor de cànnabis per tal d'entendre el seu comportament així com la seva interacció. D'aquesta manera s'han fet algunes hipòtesis sobre el pes d'algunes variables sobre el consum de cànnabis. Aquest estudi s'ha fet aplicant anàlisi de correspondències múltiple sobre les variables purament categòriques que esperàvem que tinguessin (o no) efectes significatius.

El primer que hem volgut visualitzar és l'ús de cànnabis en funció de l'edat, l'educació i el gènere de les persones consultades. De l'anàlisi es destaca la forta associació entre els homes i l'ús de cànnabis. De la mateixa manera hi ha una associació de consum amb la gent jove i en especial el aquells que abandonen els estudis. D'altra banda les dones i la gent més gran, així com aquelles persones amb formació superior estan associades a no utilitzar cànnabis. Aquests resultats eren força esperables però reafirmen una correlació significativa entre determinats grups poblacionals i l'ús de cànnabis la qual cosa ens pot aportar una bona potència predictiva.

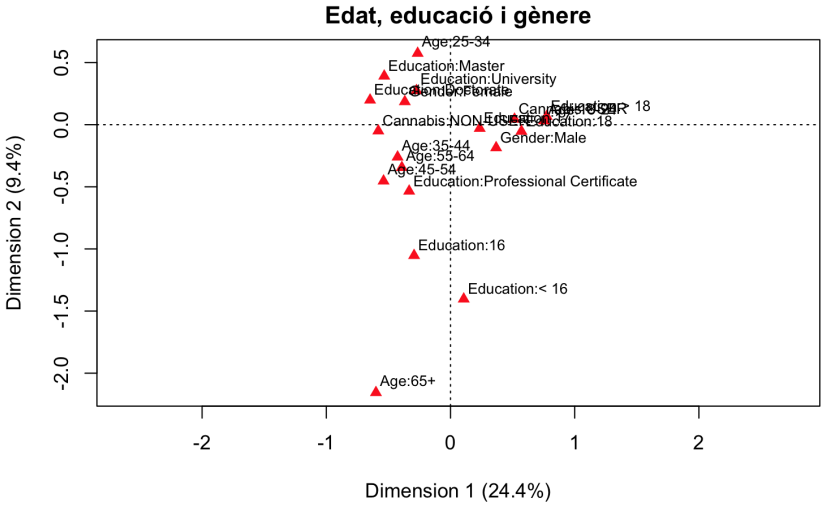

Figura 1: MCA edat, educació i gènere

També s'ha volgut estudiar la influència de drogues com l'alcohol i la nicotina sobre el consum de cànnabis. És clar que tant el consum de nicotina com alcohol tenen una forta relació amb el consum de cànnabis, de la mateixa manera que no consumir alcohol ni nicotina són actituds associades al no ús.

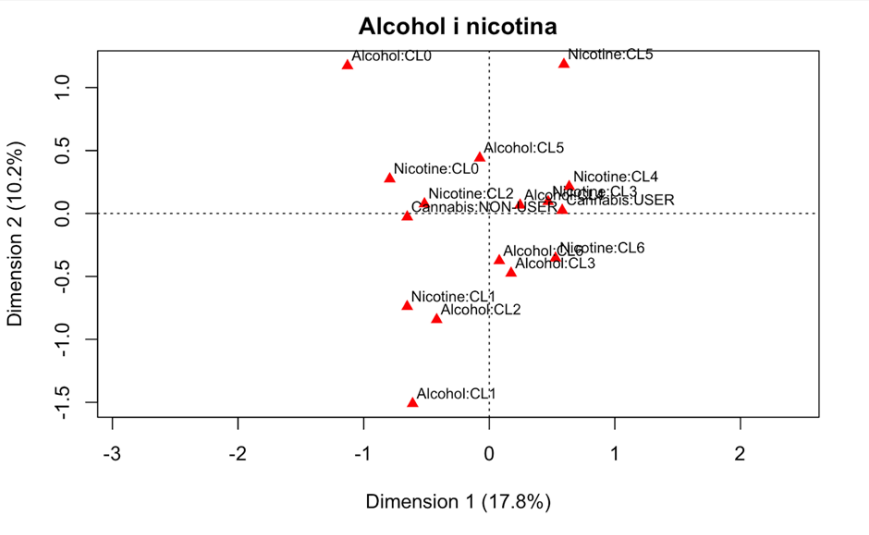

Figura 2: MCA alcohol i nicotina

Podem veure una relació de l'ús de cànnabis amb el país d'origen dels individus enquestats.

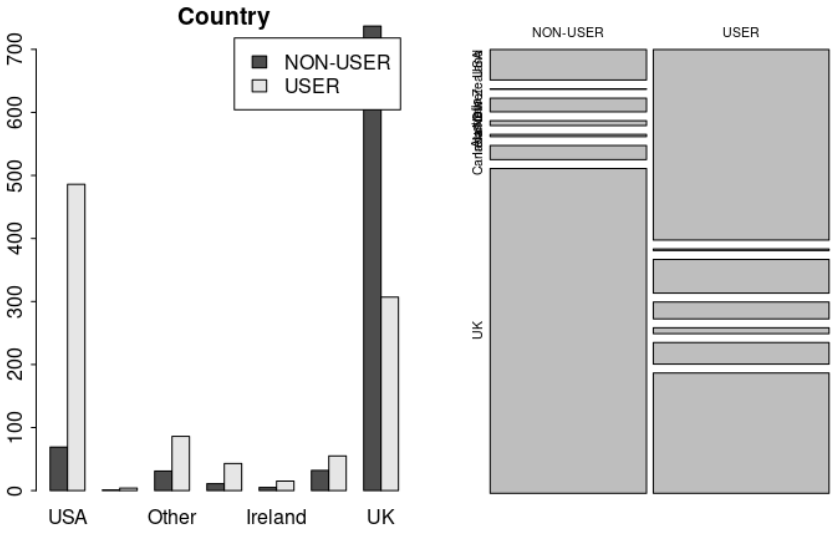

Figura 3: Barplot Consum i País d'origen

La representació de la taula de freqüències ensenya que hi ha una clara diferència entre usuaris de cànnabis a tots els països. Aquest fet ens indica que és un factor determinant en l'ús de la droga en la població. Es destaca sobretot que gran part dels usuaris britànics no en són consumidors, al contrari que els estatunidencs que majoritàriament ho són. El canvi de tendència indica que la variable és significativa en la selecció de models de predicció.

Un cop estudiades les dades demogràfiques, falta comentar la influència dels tests de personalitat sobre l'ús de cànnabis. Per fer això es durà a terme un test de Wilcox (apartat 1.5 del document annex) per a cada un de les variables de personalitat per tal d'estudiar la diferència de mitjanes entre usuaris i no usuaris de cànnabis. L'hipòtesi nul·la del test és doncs la igualtat de mitjanes d'ambdós grups. Per al correcte funcionament del test hem de garantir la normalitat de les variables estudiades. En el document annex apartat 1.5.1 podem comprovar mitjançant els *qqplots* que les variables aproximen suficientment bé la normal com per a donar per certa l'assumpció. A través del test de diferència de mitjanes podem detectar aquelles variables que no resulten ser significatives en diferenciar les dues classes i que per tant podrem apartar de l'estudi i de la proposició d'un model de predicció. A més, podem fer una representació gràfica de les distribucions de les variables explicatives de personalitat amb *barplots* que ens ajudin a visualitzar els resultats del test de Wilcox.

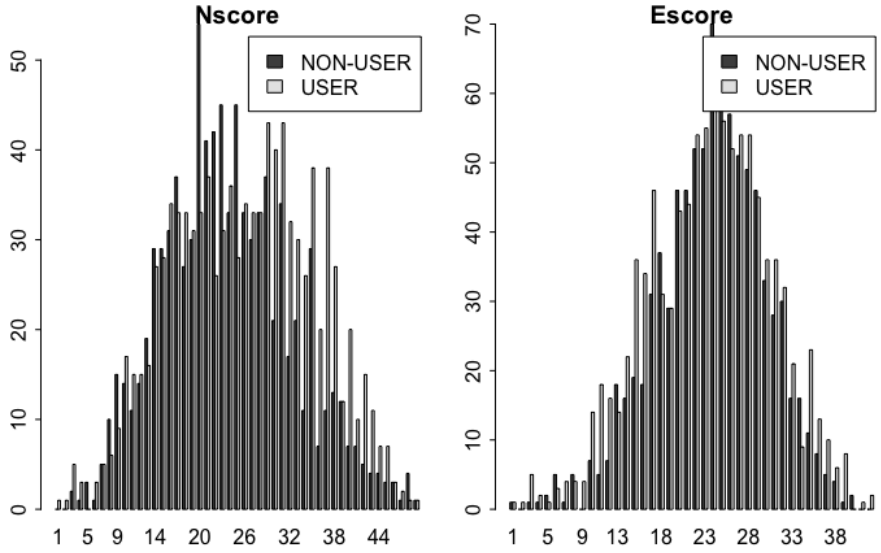

Figura 4: Histograma Nscore i Escore

Podem observar que les persones amb un grau alt de neurotisme (Nscore) presenten una major propensió a ser consumidores de cànnabis. Si analitzem la gràfica d'extraversió (Escore) veiem una forma quasi igual en ambdues classes i ens duu a pensar que la diferència podria no ser significativa. Mitjançant el test de Wilcox afirmem que no es pot rebutjar la hipòtesis nul·la i per tant no hi ha una diferència significativa entre les dues classes respecte el nivell d'extraversió de l'individu. A arrel d'això es decideix treure Escore del conjunt de variables explicatives a l'hora de desenvolupar el model de predicció.

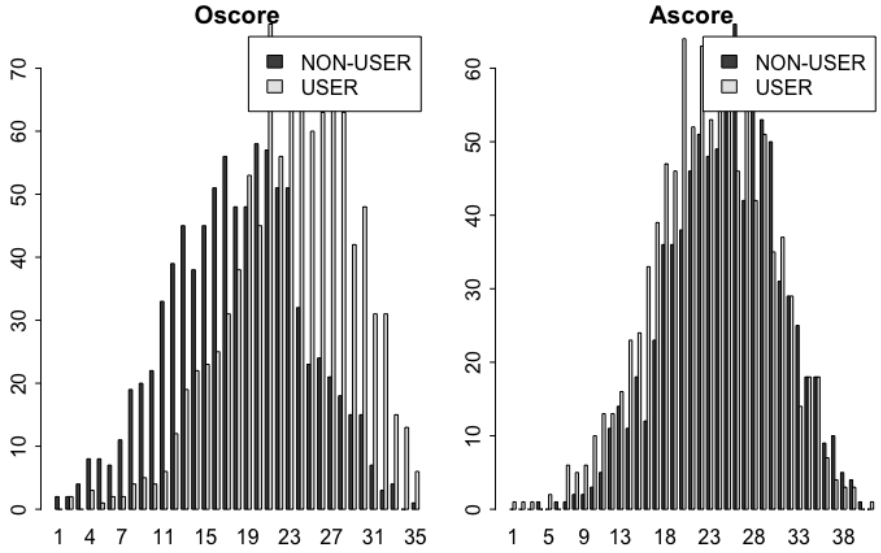

Figura 5: Histograma Oscore i Ascore

En analitzar el grau d'obertura a noves experiències dels usuaris i no usuaris de cànnabis observem en el primer grup una puntuació significativament més elevada fet que corroborem amb el test de diferències de mitjanes. Per l'altra banda els usuaris de cànnabis projecten lleugerament més baix en el test d'agrat respecte

aquells que no en consumeixen habitualment. La diferència de mitjanes per l'Ascore surt significativa i per tant la variable serà útil a l'hora de desenvolupar el model.

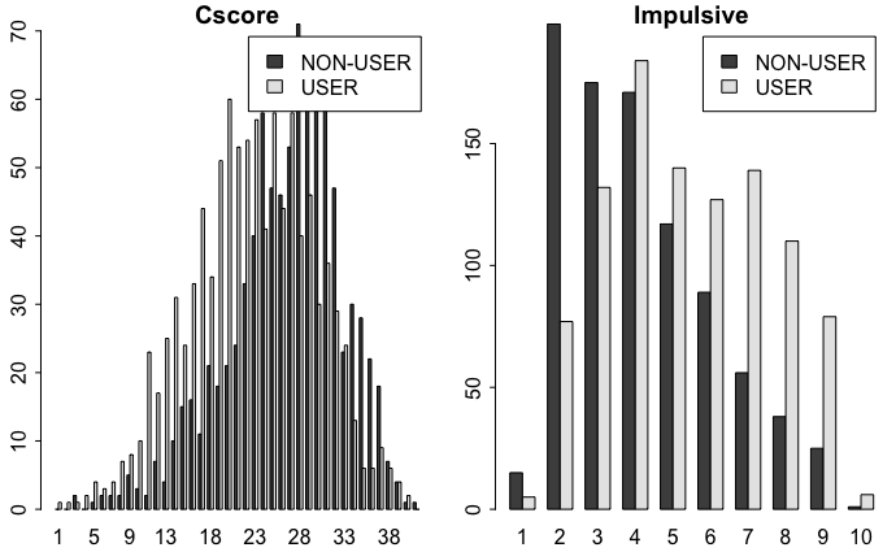

Figura 6: Histograma Cscore i Impulsivitat

Els usuaris que consumeixen obtenen, generalment, valors més baixos en el test de consistència (Cscore) que aquells que no en són consumidors. La diferència és clarament significativa. A més, els usuaris de cànnabis tenen, clarament, valors més alts d'impulsivitat.

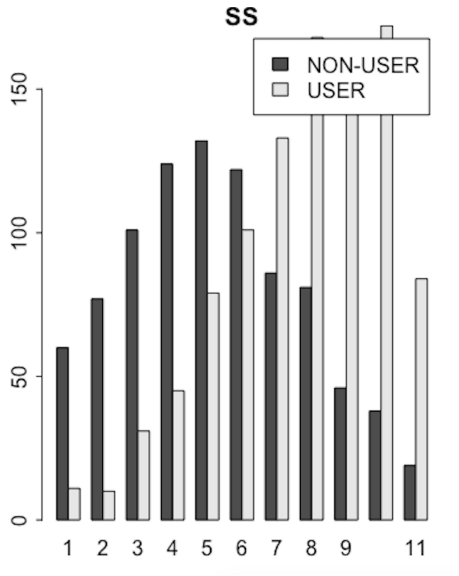

Figura 7: Histograma SS

Del test SS (Exploració de sensacions) podem observar valors generalment superiors en els usuaris de cànnabis. La diferència entre les dues classes és significativa.

A partir d'aquestes mesures podem dibuixar un perfil característic de les persones consumidores de cànnabis. De manera general, la població usuària de la droga

tendeix a ser homes joves sense estudis acabats. Mostren comportaments més impulsius, altament oberts a les noves experiències i, per tant, a explorar noves sensacions. També podem observar que tenen un grau de consistència menor, és a dir, gent menys ordenada i sense conviccions fortes. De la mateixa manera mostren, tot i que en menor mesura, un major grau de neurotisme i agrat menor que la població no consumidora.

4 Entrenament

4.1 Separació de les dades en training i test

Per a l'estudi dels diversos models que s'aplicaran a la base de dades s'ha separat aquesta en dos conjunts. D'una banda, el primer inclou un 75% de les observacions i es farà servir per escollir el millor model fent servir validació creuada. Aquest conjunt rep el nom de *training set*. D'altra banda, el segon conté el 25% de dades restants i es farà servir per avaluar els models plantejats. Aquest conjunt de test permet veure la capacitat que tenen els models per generalitzar, observant el seu comportament amb dades que no han vist mai. D'aquesta manera, amb el conjunt d'entrenament es buscarà quin dels models plantejats funciona millor amb la nostra base de dades.

4.2 Models per a l'entrenament

Donada la naturalesa de la nostra base de dades que barreja un conjunt de variables explicatives categòriques i un altre que pren valors en els naturals, podem aplicar un gran ventall de model predictors. En aquesta pràctica hem implementat GLMs amb diferents *link functions*, Naive Bayes, una xarxa neuronal amb una única capa oculta i un Random Forest. A més aprofitant la codificació amb la que venia la base de dades hem utilitzat QDA i LDA. Tots aquests models i, com s'explica al punt 4.3, s'han validat mitjançant un *10-times 10-fold cross validation* per discernir quin és el model òptim.

Dos dels models implementats a la pràctica, la xarxa neuronal i el bosc aleatori requereixen paràmetres previs a l'entrenament que modifiquen el comportament del model predictiu. Per escollir aquests hiperparàmetres apliquem *10-times 10-fold CV* sobre el conjunt d'entrenament. Aquest mètode consisteix en fer la mitjana de l'error de 100 entrenaments del model amb cadascun dels possibles valors que considerem per l'hiperparàmetre. D'aquesta manera, a partir d'un mostreig aleatori de les dades de *training* entrenem i validem el model amb una separació de les dades de 90% i 10% respectivament. Aquest procés el repetim 10 vegades i triem el valor que resulti en el menor error.

4.2.1 GLM

Donat que tenim una barreja de dades categòriques i naturals, una elecció clara ha sigut la implementació de models lineals generalitzats. En aquest cas, ja que es tracta d'un problema de classificació, hem utilitzat tres GLM amb tres *link functions* diferents: *logit*, *probit* i *log-log* i escollim amb *CV* la millor funció d'enllaç.

4.2.2 Naive Bayes

Una altra elecció clara de model és Naive Bayes, ja que tenim un gran conjunt de variables categòriques. El model s'entrena amb la funció de R *naive_bayes* la qual per a les variables numèriques fa una separació per quantils.

4.2.3 Xarxa neuronal

Implementem una xarxa neuronal amb una única capa oculta mitjançant la funció del paquet *caret* de R *nnet*. Abans de fer l'entrenament del model, hem d'establir el nombre de neurones a la capa oculta de la xarxa i el *decay*. A l'hora de treballar amb xarxes neuronals sabem que a mesura que augmentem el nombre de neurones, també augmenta l'ajust sobre les dades d'entrenament. D'aquesta manera molt poques neurones suposaria un $underfit$ de les dades, que resulta en errors d'entrenament i test elevats. En canvi un nombre massa elevat d'aquestes suposaria error petit d'entrenament i un error elevat de test, fenomen conegut com $overfitting$. Per altra banda, el *decay* ens permet afegir un biaix al model que evita la sobreparametrització del mateix. Haurem d'ajustar, aleshores, tots dos paràmetres per aconseguir un model equilibrat que no sobreajusti sobre el conjunt d'entrenament ni produeixi $underfitting$.

Hi ha moltes maneres de trobar el nombre de neurones i *decay* òptim per una xarxa, però la més segura és primer buscar un número de neurones que provoqui *overfit* i a continuació trobar el *decay* que li correspongui. Tot això es farà com s'ha fet fins ara, amb *10-times 10-fold CV*.

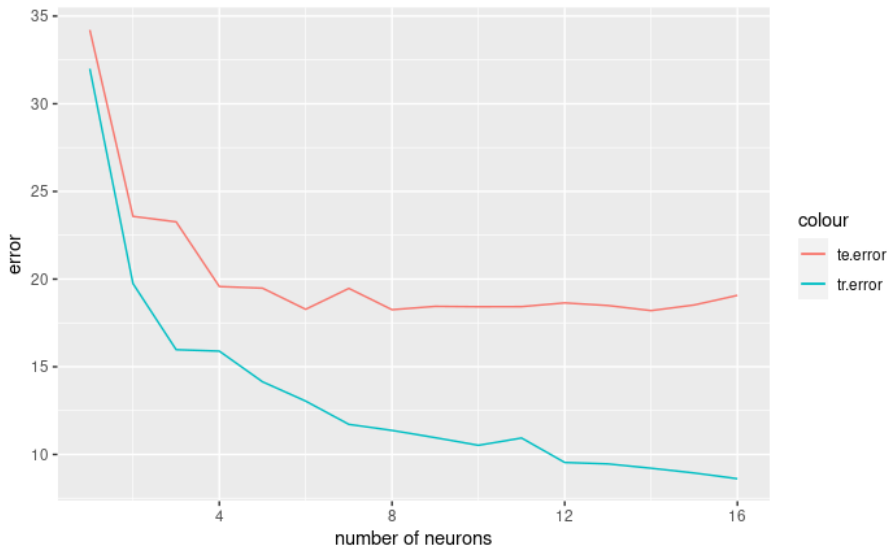

Figura 8: Error en funció del nombre de neurones

D'aquesta manera, el nombre de neurones més petit amb el que trobem $overfitting$ és de 10, ja que observem que a partir de les, aproximadament 8-9 neurones, el error de test comença a augmentar mentre que el error d'entrenament segueix disminuint. Un cop escollit el nombre de neurones que comet sobreajust podem, mitjançant *cross-validation*, trobar el *decay* que ens retorna un error mínim, en aquest cas 2.5.

4.2.4 Random Forests

En *Random Forests* hem d'ajustar el nombre d'arbres que ha de tenir el nostre *ensemble*. Un nombre alt d'arbres no provoca $overfitting$, sinó que a mesura que augmentem el nombre d'arbres ens apropem a l'error mínim, que succeeix quan el nombre d'arbres tendeix a infinit. Tot i així, aquesta convergència no és lineal, sinó que al principi l'error disminueix exponencialment en funció del nombre d'arbres i quan arriba a una determinada mida la millora obtinguda amb l'augment de la complexitat del bosc no és gaire significativa. D'aquesta manera el comportament observat seria el següent:

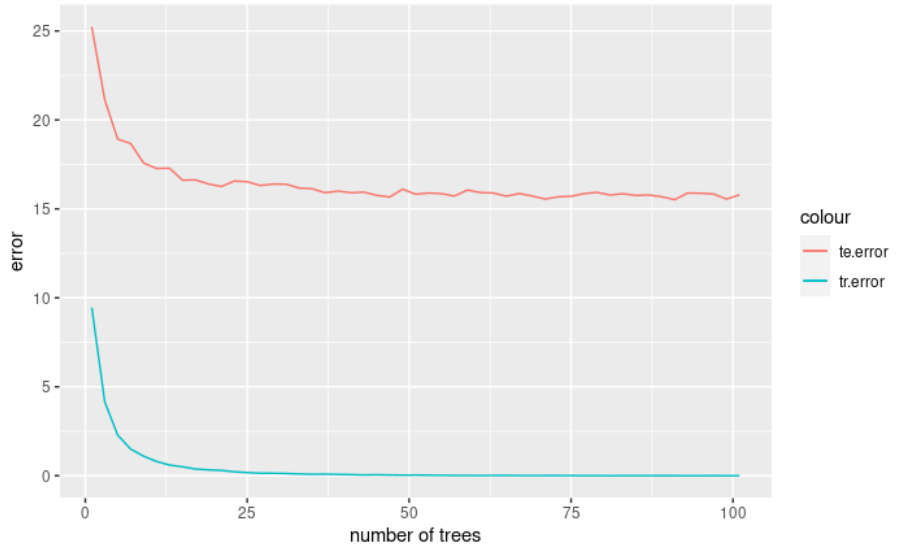

Figura 9: Error en funció del nombre d'arbres

S'observa que a partir d'una mida del bosc s'estabilitza l'error i la millora en funció del nombre d'arbres ja no és significativa. Per això triem un nombre d'arbres suficientment gran per obtenir un error petit, però no molt gran com per augmentar massa el cost computacional sense millorar-ne el resultat. En aquest cas, observem que en, més o menys, 50 arbres l'error s'estabilitza i l'augment del nombre d'arbres no provoca una disminució significativa en l'error de predicció.

4.2.5 Anàlisi discriminant

La nostra base de dades és de variables categòriques i factors ordinals, però aprofitant la codificació a reals que aproximen la distribució normal podem implementar dos mètodes d'anàlisi discriminant: LDA i QDA. Per tal de poder fer servir aquests anàlisis ens calen variables explicatives que compleixin normalitat multivariada. Un cop aplicada la transformació cap a normalitat que ens ha donat la base de dades podem garantir la normalitat individual de cada variable. Per saber si podem assegurar també la normalitat multivariada de tot el conjunt de variables hem de fer un test equivalent al *qqplot* però en múltiples dimensions, que en el nostre cas utilitza la distància de Mahanalobis.

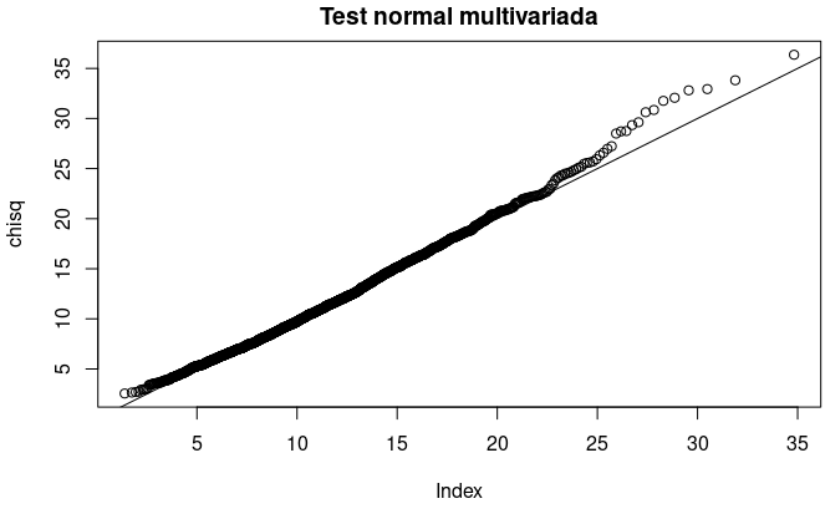

Figura 10: Test de normalitat multivariant

Veiem que els punts s'ajusten prou bé a la línia de la normal; la hipòtesi de normalitat multivariada es compleix, i per tant podem garantir un resultat coherent dels mètodes de creació de models mitjançant l'anàlisi discriminant. En fem tant l'anàlisi lineal (on considerem igualtat de covariàncies, podem veure el plot de LDA al document annex apartat 1.6.1) com l'anàlisi quadràtic (on les covariàncies difereixen).

4.3 Validació dels models

Un cop plantejats els model per a l'entrenament, hem de discernir quin aconsegueix modelar millor la nostra base de dades. De la mateixa manera que quan s'han buscat els hiperparàmetres, el millor model s'ha determinat amb un *10-times 10-fold cross validation* on s'ajusta cada model amb un 90% de les dades d'entrenament i es valida sobre el 10% de les dades d'entrenament restants. Aquest procés es repeteix 100 cops (10 permutacions de la base de dades i 10 folds per a cada permutació). Això ens permet obtenir una estimació robusta de l'error de cada model i escollir el més adient. Es pot veure com s'ha fet amb més detall amb el codi a la secció 2.8 del document annex. Els resultats dels errors de validació obtinguts per a cada model són els següents:

QDA	*LDA*	*GLM logit*	*GLM probit*	*GLM cloglog*	*Naïve Bayes*	*ANN*	*RF*
19.57	18.46	15.95	15.97	17.12	17.31	16.04	15.91

A primera vista, amb els models obtinguts, podem observar que tots els models proporcionen errors entre el 15% i 20% En els models d'anàlisi discriminant (QDA i LDA) on utilitzàvem la codificació original de les dades a reals s'ha obtingut un error significativament major 19.57 per a QDA i 18.45 per a LDA. Això ens dona l'intuïció que les matrius de covariàncies d'ambdós grups podrien ser equivalents. Per una altra banda, els tres GLM han funcionat prou bé. Aquells on s'ha utilitzat una funció d'enllaç simètrica han donat menor error que el GLM amb funció d'enllaç *log-log*. Naive Bayes també prediu prou bé sobre el conjunt de validació però mostra un error d'un 1% més gran que els millors models, que se situen a la vora del 16%.

Aquest tant per cent d'error el veiem amb la xarxa neuronal amb una única capa oculta i amb el Random Forest.

Així doncs, hem obtingut 4 models que aconsegueixen un error mínim molt semblant a prop del 16% amb una desviació d'aproximadament un decimal. En són els models lineals generalitzats binomials amb funció d'enllaç simètrica, la nnet i el Random Forest. Per tant, escollim aquests models com els millors per a la nostra base de dades. Entrenarem, per tant, tres models finals: un amb la nnet, un amb el Random Forest i un altre amb els GLM *logit* ja que per la construcció dels models lineals generalitzats, el model final quedarà molt semblant amb qualsevol dels dos *link functions* simètrics.

5 Test

Un cop escollits el models òptims per a la base de dades, podem entrenar els models finals i calcular l'error final sobre el conjunt de test que vam crear a l'inici. A la següent taula es poden observar els errors d'entrenament i test obtinguts per a cada model tant en el conjunt de *training* com en el de *test*.

Model	tr. error	te. error
GLM logit	13.93	15.50
ANN	13.86	16.35
RF	0.07	17.20

Els models GLM tenen bona actuació, utililitzant la funció d'enllaç *logit* obtenim un error de test del 15.5%. Les taules de contingència de comparació entre els valors originals i els predits ensenyen que els falsos negatius (32) i els falsos positius (41) estan equilibrats entre les dues classes. Els valors d'error de *training* i de test són similars, per tant no observem *overfitting* del model.

El mètode de la xarxa neuronal ens proporciona un model amb un error de test del 16.3%. Aquest es troba en equilibri entre les dues classes estudiades, hi observem 41 falsos positius i 36 falsos negatius, on podem considerar l'error equilibrat entre les classes. Tampoc es detecta *overfitting* al model ja que els valors d'error de test i *training* són prou semblants.

En el cas del *Random Forest* observem que per les dades de test l'error és del 17%. Aquest error està bastant balancejat entre ambdues classes, ja que s'observen 39 falsos positius i 42 falsos negatius. Pel que fa a les dades d'entrenament, veiem que l'error és gairabé 0, amb només una observació mal classificada. Com s'ha explicat anteriorment, això no és resultat de sobreajust, sinó que el nombre d'arbres escollit és prou gran com per apropar-se a l'òptim per les dades d'entrenament.

6 Conclusions

Hem obtingut un model prou eficient en predir el consum de cànnabis a la població estudiada: l'error de predicció és vora el 15% quan s'aplica a les mostres separades per fer el test. Això ens duria a pensar que és un bon model, atenent que l'objectiu

de l'estudi és la classificació segons si l'individu és consumidora o no de cànnabis. Com que el cost dels falsos positius i els falsos negatius és igual, i el fet que el model s'equivoqui és comprensible, ja que estem tractant temes socials, podem considerar que el model obté bones prediccions.

Malgrat obtenir un error relativament baix i per tant semblar haver trobat un model prou precís per generalitzar a la població general, no podem garantir que, en aplicar el model a la població general, es cometi el mateix error. Hem comentat en la descripció del conjunt de dades que l'obtenció dels resultats de l'enquesta va ser realitzada amb el mètode *Snowball*. Aquest mètode de captació de dades pot donar lloc a prediccions esbiaixades ja que no podem assegurar-nos que tenim una mostra poblacional prou representativa ni de la correctesa de les probabilitats a priori dels models.

Veiem indicacions d'aquestes possibles inconsistències en diversos moments com ara a la variable *country* que mostra que la majoria de respostes rebudes procedien o bé dels Estats Units o del Regne Unit. Veiem que la majoria dels habitants d'EUA són consumidors i la majoria dels britànics no ho són. Encara que aquest paradigma podria produir-se, el fet que l'enquesta s'escampa a través de cercles de coneixença ens duu a pensar que podríem haver rebut resultats extrems degut a la falta de representació de població a qui no ha arribat el qüestionari. També observem una presència superior de població juvenil, l'estil de vida de qui, a priori, sembla tendir més al consum habitual de cànnabis. Aquest fet pot haver-se produït també pel fet que, l'enquesta era *online* i, per tant, no aconsegueix representar correctament a la població més gran.

Així doncs, amb aquest projecte hem aconseguit aprofundir en el desenvolupament de mètodes de classificació i estudiar com afecten en els resultats aspectes tan variats com són la captació i tractament de dades o la influència dels hiperparàmetres en els models.

Referències

[1] E. M. Mirkes V. Egan E. Fehrman, A. K. Muhammad and A. N. Gorban. *Drug Consumption Data Set*. UCI Machine Learning Repository, 2015.

[2] E. M. Mirkes V. Egan E. Fehrman, A. K. Muhammad and A. N. Gorban. *The Five Factor Model of personality and evaluation of drug consumption risk*. arXiv, 2015.

[3] E. Fehrman and V. Egan. *Drug consumption*. ICPSR, 2012.