Caracterització de perfils acadèmics

Projecte de Machine Learning (AA1)

Marina Rosell Murillo

Pau Lozano García

Patricia Cabot Álvarez

Grau en Ciència i Enginyeria de Dades

Juny de 2020

Prof. Lluís Antonio Belanche Muñoz

Prof. Marta Arias Vicente

Índex

1. Introducció

Aquest estudi busca examinar els diferents criteris que determinen el rendiment escolar del jovent a l'institut, en base a l'anàlisi de múltiples factors.

El propòsit del treball és elaborar i comparar un seguit de models predictius que posteriorment permetin extreure conclusions sobre els elements que incideixen en la qualificació dels estudiants.

La base de dades ha sigut obtinguda d'un estudi realitzat anteriorment per la Universitat de Minho, Portugal[1]. La seva descripció s'explica amb més detall en el següent apartat.

La descripció proporcionada de les dades a l'estudi explica que aborda l'èxit dels estudiants a l'educació secundària de dues escoles portugueses. Les dades característiques inclouen les qualificacions dels estudiants, com a variables resposta, i factors demogràfics, socials i relacionats amb l'escola. Més endavant es proporciona una breu explicació de cada variable.

La informació va estar recollida utilitzant qüestionaris i informes cada escoles. El conjunt de dades consisteix en dos bases de dades amb la informació de dues assignatures diferents: matemàtiques i portugués. Cada asignatura consta de tres variables numèriques diferents, que coincideixen amb les notes del primer i segon trimestre i amb la nota final.

2. Descripció de treball previ

Com hem especificat anteriorment, l'estudi està dirigit a la predicció de les qualificacions dels estudiants d'educació secundària de dues assignatures, matemàtiques i portugués, utilitzant notes escolars antigues (del primer i el segon trimestre) i dades demogràfiques, socials i relacionades amb l'escola.

Aborda la investigació amb quatre mètodes diferents: arbres de decisió (DT), boscos aleatoris (RF), xarxes neuronals (NN) i *Support Vector Machine* (SVM). A més, per cada mètode, ha estat explorat amb diferents seleccions de variables d'entrada.

Els resultats obtinguts revelen que és possible aconseguir una predicció d'alta exactitud, donades les notes del primer i el segon trimestre. L'anàlisi elaborat que millors prediccions dona demostra que els factors demogràfics, els socials i els relacionats amb l'escola són molt rellevants també, a part de les qualificacions de l'alumnat.

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. http://archive.ics.uci.edu/ml/datasets/Student+Performance

3. Procés d'exploració de les dades

3.1 Pre-processament

Primer, disposem a fer una breu explicació de les variables proporcionades per la base dades. Aquesta descripció es troba a *Annex Taula 1*.

3.3.1 Anotacions addicionals sobre les variables:

- Notes finals és la variable que volem predir.
- Les següents variables són compartides entre assignatures:
    - Escola, sexe, edat, direcció, tamany de la família, situació dels pares, educació de la mare, educació del pare, feina de la mare, feina del pare, raó, tutor, temps de viatge, ajuda familiar, guarderia, superior, internet, relació romàntica, relació familiar, temps lliure, quedar, alcohol entre setmana, alcohol caps de setmana i salut.
- Les següents variables són pròpies de cada assignatura:
    - Temps d'estudi, suspeses, ajuda escolar, pagament, activitats, absències, notes 1r trimestre, notes 2n trimestre i notes finals.

Quan analitzem inicialment les dades sense processar, tenim:

Base de dades de matemàtiques	33 columnes (variables)	395 files (estudiants)
Base de dades de portugués	33 columnes (variables)	695 files (estudiants)

No s'observa cap anomalia a les bases de dades ni cap dada mancant.

Per poder realitzar càlculs numèrics sobre les bases de dades i modelització, hem fet una binarització de les variables categòriques (*one hot encoding*).

3.2 Extracció i selecció d'atributs

Durant la primera investigació i els primers càlculs que hem fet servir per familiaritzar-nos amb les dades, ens hem adonat de seguida que n'hi havia massa. No obstant, conforme hem anat avançant en el nostre anàlisi ens hem adonat de que en realitat totes tenen coses a dir i donen informació important. Així doncs, per facilitar la visualització, i afavorir en la velocitat dels càlculs, hem decidit que separaríem les dades per temàtica i que analitzaríem les dades des de diferents punts de vista. Les separacions han estat les següents:

1. Família:
"famsize", "Pstatus", "Medu", "Fedu", "Fjob", "Mjob", "guardian", "famsup", "famrel"
Aquesta separació recull les variables que considerem que poden influir en les notes de l'estudiant que estan relacionades amb la seva família.

2. Personal:

"sex", "age", "health", "nursery"
Són variables que ens caracteritzen cada estudiant en funció de les seves dades personals com podrien ser el sexe, l'edat, el seu estat de salut i si va anar a la guarderia quan era petit.

3. Environment:
"address", "traveltime", "internet"
L'ambient de l'estudiant i el seu entorn son claus en el seu desenvolupament acadèmic, cal diferenciar si viu en una zona urbana o rural, si el seu temps de desplaçament és llarg o si disposa de connexió a internet.

4. Social:
"activities", "romantic", "freetime", "goout", "Dalc", "Walc"
Un altre factor molt interessant per analitzar el rendiment acadèmic pot ser l'ús del temps extraescolar. Per tant cal veure si l'estudiant té una relació romàntica amb alguna altre persona, si realitza activitats extraescolars, la quantitat del seu temps lliure, si surt gaire sovint amb els seus amics, i el seu consum d'alcohol entre setmana i els caps de setmana.

5. School:
"school", "reason", "schoolsup"
Les característiques de l'escola influeixen enormement en l'estudiant, i en alguns casos ens pot ajudar a predir quina és la seva situació. Avaluem en quina de les dues escoles de l'estudi es troba l'estudiant, la raó per la qual està en aquesta escola (distància de casa, prestigi etc.) i si la escola ofereix a l'estudiant classes extra de reforç.

6. Student:
"studytime", "failures", "higher", "absences", "paid"
La motivació i l'expedient acadèmic de l'estudiant és una font primordial de dades per predir el seu desenvolupament. Es pren una mesura del seu temps d'estudi, la quantitat d'exàmens que ha suspès, si pretén estudiar més enllà de secundaria, la quantitat d'absències durant el curs i si paga per una acadèmia o classes particulars.

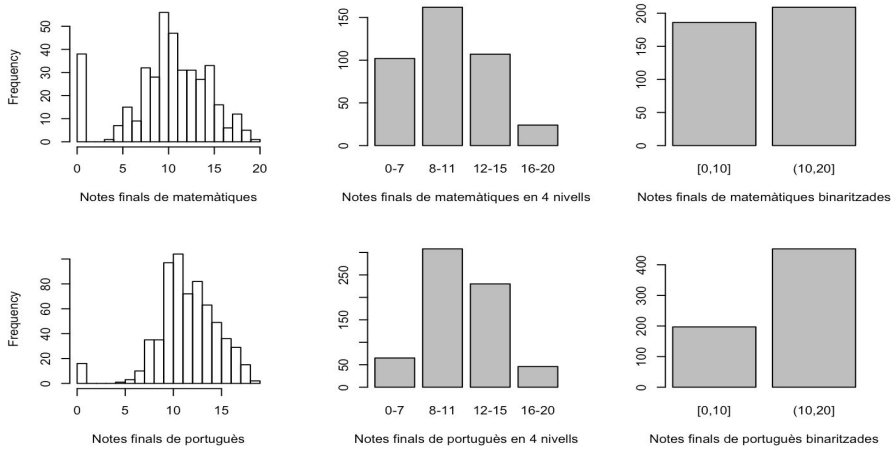
7. Grades:
"G1", "G2"
Per últim, la millor font d'informació que disposem per predir com li anirà a un estudiant el trimestre d'una assignatura, és saber com ha estat el seu rendiment en els trimestres anteriors. G1 i G2 són les notes corresponents a l'assignatura en els dos primers trimestres del curs.
Hem decidit modelar les dades de 3 formes diferents segons la variable resposta G3:

- Classificació binària (aprovat/suspès)
- Classificació en 4 nivells segons la nota (0-7, 8-11, 12-15, 16-20)

En aquesta classificació hem volgut diferenciar els 4 grups com: suspensos lluny de l'aprovat, raspant l'aprovat, aprovat tranquil i nota molt alta.

- Regressió amb un output numèric entre 0 i 20 (notes mínima i màxima)

Histogrames corresponents a les tres modelitzacions:



3.3 Visualització

Podem començar tenint una noció general del comportament de les nostres dades i de quina manera influeixen en el desenvolupament de l'estudiant.
Aquí, per facilitar la visualització i simplificar els gràfics, utilitzarem la separació per temàtiques de les variables, explicada a 3.2.

3.3.1 Anàlisi de components principals

Els següents gràfics diferenciaran les observacions que han aprovat (color blau cel) de les observacions que han suspès (en taronja). Com és d'esperar, les observacions es distribueixen d'una forma que diferencia dos grups (tot i que es veu certa barreja perquè només es representen els dos primers components principals i no són capaços d'explicar-nos tota la variància de les dades).

Les fletxes, les podem interpretar com el comportament que tenen les variables respecte de les dades. Si una fletxa apunta cap al grup dels aprovats, interpretem que aquesta variable quan té un valor alt "afavoreix" a que l'estudiant tingui notes per sobre de l'aprovat. Contràriament, si una fletxa apunta cap al grup dels supesos, associem els valors alts per aquesta variable amb les notes per sota de l'aprovat.

Comencem amb la separació de les variables relacionades amb la familia:
Annex Imatges 1, 2
Contràriament al que podríem pensar, no és gaire evident que l'educació dels pares influeix en el desenvolupament acadèmic de l'estudiant.
Si ens fixem en la resta de variables, veiem com la mida de la família sorprenentment no juga un paper rellevant en el rendiment acadèmic. L'estat civil dels pares, té una forta rellevància en l'assignatura de matemàtiques (en el codi R es pot fer zoom i s'aprecia millor que en la foto que s'ha adjuntat) essent favorable a les famílies on els pares NO viuen junts (de fet, el divorci dels pares ens ha aparegut molt associat al grup que treu millors notes en un altre anàlisi per categories que hem realitzat). Per a l'assignatura de portuguès en canvi l'estat civil dels pares no sembla jugar un paper rellevant.

Per últim, com era d'esperar, el fet que l'estudiant senti el suport dels seus pares s'associa molt al grup dels aprovats.

Separació de variables relacionades amb la professió dels pares de l'alumne:

Annex Imatges 3, 4

Aquest gràfic és molt ambigu i treu poc en clar, ja que la majoria de les variables només tendeixen lleugerament cap a alguna de les bandes.

En tot cas, sí que sembla que si els pares es dediquen a les ciències de la salut, l'estudiant tendeix a treure millors notes. Contràriament, a les famílies on almenys un membre de la familia es dedica a les tasques de casa (on també es contempla l'atur) l'estudiant tendeix a treure males notes.

Podem destacar les professions relacionades amb l'ensenyança, que tendeixen molt lleugerament al grup dels aprovats.

Per a la resta de professions, la relació respecte a les notes de l'alumne no sembla tenir un paper especialment important.

Separació de variables relacionades amb l'àmbit personal de l'alumne:

Annex Imatges 5, 6

Podem deduir que aspectes com haver anat a la llar d'infants té una influència positiva a les notes dels alumnes bastant marcada, tant en matemàtiques com en portuguès, ja que la fletxa de la variable corresponent apunta en direcció als grups dels aprovats.

Per altra banda, veiem que, com més edat tenen els estudiants, les notes que treuen són més baixes, probablement degut a que aquests estudiants que tenen més anys siguin repetidors.

Aspectes com el sexe i la salut actual de l'alumne no semblen ser gaire influents a la seva nota final, ja que les direccions de les fletxes van entremig els dos grups, com a dada extra podem observar que els homes tendeixen a tenir millor salut que les dones.

Separació de variables relacionades amb l'entorn de l'alumne:

Annex Imatges 7, 8

Poques coses a comentar aquí, i totes molt esperables.

La variable adreça és binaria (rural o urbana) i s'observa com el fet de viure a la ciutat afavoreix a aprovar. Tenir accés a internet també ho associem a treure millors notes. Per últim, l'increment del temps de desplaçament resulta en un impacte negatiu en el rendiment acadèmic.

Separació de variables relacionades amb l'àmbit social de l'alumne:

Annex Imatges 9, 10

Pel que respecta a la vida social de l'alumne, en general no hi ha associacions fortes entre variables i les notes de l'estudiant. Però cal matisar algunes coses.

Primer, i la que és més evident: els estudiants que tenen relacions romàntiques treuen pitjors notes. Ho atribuïm a que aquests estudiants disposen en general de menys temps d'estudi.

Les activitats extraescolars tendeixen lleugerament al grup dels aprovats, mentres que el consum d'alcohol influeix negativament però no d'una forma especialment marcada. Això, com observarem més endavant, és degut a que la majoria dels estudiants reconeixen

consumir alcohol de forma moderada (al voltant dels nivells 2 i 3) siguin quines siguin les seves notes. No obstant, sí que hem trobat una forta relació entre el consum d'alcohol mínim i el grup d'estudiants que treuen millors notes, i també hem trobat molta relació entre la consumició alta d'alcohol i treure molt males notes. Les variables "freetime" i "goout" es comporten de forma semblant al consum d'alcohol.

Variables relacionades amb l'àmbit de l'escola de l'alumne:

Annex Imatges 11, 12.

Veiem que els estudiants que van escollir l'escola per la seva reputació treuen millors notes tant a matemàtiques com a portuguès, tot i que la relació és més marcada a l'assignatura portuguès. En canvi, aquells alumnes que van escollir l'escola a la que estudien per preferència d'un curs concret, treuen pitjors notes. Aquells que van escollir l'escola per altres motius no semblen marcar cap tendència els seus resultats.

Una altra variable que sembla afectar directament als bons resultats acadèmics és rebre ajuda educativa extra per part de l'escola, això té sentit i és raonable, ja que aquest tipus d'ajudes estan pensades per ajudar als alumnes a millorar els seus resultats, i semblen ser efectives.

Sorprenentment, la variable escola també influeix sobre les notes dels alumnes, en ambós assignatures els alumnes que són de l'escola Moushino da Silveira obtenen millors resultat que els de l'escola Gabriel Pereira.

Variables relacionades amb els hàbits i objectius de l'estudiant:

Annex Imatges 13, 14.

Aquí la interpretació és molt clara i molt ràpida: els estudiants que estudien més temps són els que tenen millors notes.

Destaca també la variable "higher", que ens indica si l'estudiant té intenció de seguir estudiant en anys posteriors, els alumnes que ho afirmen solen treure millors notes.

Contràriament, els estudiants que acumulen moltes absències a classe tendeixen a treure pitjors notes i aquells que han suspès exàmens anteriors, també.

Per últim, no sembla que les classes particulars tinguin un paper gaire rellevant en les notes finals de l'alumne.

Variables relacionades amb les notes de les avaluacions anteriors de l'estudiant:

Annex Imatges 15, 16.

Clarament observem relació directa en què l'alumne hagi aprovat a les avaluacions anteriors i que les notes finals també siguin aprovades, així com també observem una relació molt forta entre els alumnes que aproven les assignatures al 1r i 2n trimestre. És a dir, si un alumne ha aprovat el primer trimestre, és molt probable que també aprovi el segon, i si és donen aquests dos casos és encara més probable que aprovi la avaluació final. Aquesta relació es manté a les dues assignatures.

3.3.2 Anàlisi discriminant lineal

Imatges Annex 17, 18, 19 i 20.

Els resultats obtinguts pel mètode d'anàlisi discriminant lineal sobre tota la base de dades mostren una clara separació entre les categories de notes per ambdues assignatures.

4. Protocol de re-mostreig i models considerats

4.1 Re-mostreig

Hem realitzat una reordenació de les bases de dades que utilitzarem en els models de manera completament aleatòria. Posteriorment, hem dividit cada una en train (¾) i test (¼). Ens assegurem que en cadascuna de les particions, tenim la mateixa proporció de cada grup de classificació de notes.
Annex Taules 2, 3.
observem a les taules que els grups estan ben representats a cada base de dades dividida.

4.1.1 Cross Validation

Fem servir aquest mètode per calcular l'error de CV corresponent de cada model considerat i d'aquesta manera tenir més informació sobre precisió de cadascun. Ho farem sobre la base de dades de train.

Farem servir un procediment iteratiu amb $k = 10$, de manera que dividirem la base de dades utilitzada (de train) en 10 parts iguals. Considerarem 9 de les parts com a base d'entrenament i la desena com a base de validació. A cada iteració, entrenarem cada model seleccionat amb la base d'entrenament i calcularem el seu error de validació.

4.2 Models considerats

A continuació, passem a explicar els diferents models considerats en aquest estudi.

- **Model GLM**. Inicialment crearem els models amb totes les variables i a continuació farem servir la funció step per reduïr el nombre de variables del model i d'aquesta forma simplificar el model. Farem servir una distribució Gaussiana al model que tracta la variable resposta (notes finals) com a numèrica (el model de regressió), i un distribució Binomial al cas que tractem la variable resposta binaritzada (aprovat/suspès), ja que hem de fer servir el logit com a funció link per tal que la resposta sigui binària. Pel cas de la nota final explicada amb quatre categories, no podem fer servir aquest model, ja que no accepta aquest format.

- **Model amb Anàlisi Discriminant Lineal**. Aquest model és molt interessant ja que ens redueix totes les variables en el nombre de grups menys una, de manera que el model resultant está molt simplificat, cosa que sempre va bé per agilitzar càlculs i reduir el tamany del data frame. Està especialment pensat per fer discriminacions d'observacions i aplicarem aquesta discriminació segons la predicció de les notes finals.

- **Model amb Naive Bayes**. Hem fet servir aquest model perquè serveix per fer discriminacions probabilístiques i ens va molt bé per predir aquesta discriminació

segons les notes finals. A diferència del model amb Anàlisi Discriminant Lineal fa servir el mètode de màxima verosimilitud i les característiques estadístiques de la distribució com les priors i posteriors.

- **Model amb K-Nearest Neighbours**. Aquest model agafa una mostra de test i mira els k veïns més propers d'aquesta en la mostra de train. La predicció que realitza es fa mitjançant votació, és a dir, es pren el grup que predomina en aquests k veïns utilitzats. En cas d'empat, es pren un dels possibles valors a l'atzar.

- **Model amb Xarxes Neuronals**. Les xarxes neuronals artificials són un mètode molt estés en el món del Machine Learning. El principi que segueixen és una aproximació a les xarxes neuronals biològiques. La idea és crear una xarxa de "neurones" que responen a un input amb un output que pot anar a una nova neurona o utilitzar-se com a resposta. Cada neurona té uns coeficients (pesos) que ponderen les dades d'entrada i una funció d'activació que decideix la resposta de la neurona.
  La selecció d'aquests pesos és un problema d'optimització, ja es prenen per tal de minimitzar l'error de sortida de la xarxa.

- **Model amb Random Forest**. Els models de Random Forest són molt útils per classificació i regressió, els quals es formen entrenant múltiples arbres de classificació. Per tant, poden aportar molts bons resultats al tipus de predicció que estem realitzant.

5. Resultats preliminars i comparació

Per realitzar l'anàlisi de la qualitat dels models sobre la predicció de les qualificacions dels estudiants, hem entrenat els models anteriors i hem dut a terme diferents anàlisi. Com a mesura de qualitat per classificació (notes en 2 i 4 classes), hem calculat el percentatge de precisió utilitzant una taula de confusió. Per altra banda, en el cas de regressió, hem calculat l'error quadràtic mitjà.

5.1 Selecció dels hiper-paràmetres dels models

A continuació, especifiquem el mètode emprat per determinar els hiper-paràmetres dels algorismes que els requereixen.

- **K-Nearest Neighbours.**
  Hem realitzat un algorisme iteratiu en què provem diferents valors per k (valors de 1 a 10), i ens quedem el que millor prediu les dades de validació. El resultat òptim per als sis casos ha resultat ser $k = 1$. Els resultats de l'algoritme es troben a *Annex Taules 4, 5, 6, 7, 8, 9*.

- **Xarxes Neuronals.**
  Hem realitzat dos tests de Cross Validation per decidir en primer lloc el número òptim de neurones a la capa interna havent fixat la regularització (decay) a zero. Un cop trobat el el valor òptim, hem fixat el nombre de neurones a la capa interna a vint (el valor que hem considerat com a màxim) i hem realitzat el Cross Validation variant el valor del decay (de zero a u). En cada cas hem escollit la combinació que ens ha donat millors resultats dels dos tests (major accuracy pels casos de classificació i menor RMSE pels casos de regressió). Els resultats obtinguts es troben a *Annex Taules 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21*. En tots els casos el millor model és el que inclou una mesura de regularització i fem servir el nombre màxim de neurones ocultes (20), en els casos de classificació amb 2 categories i regressió el valor de decay.

- **Random Forest.**
  Hem realitzat un procediment iteratiu on creem models de RF amb les dades d'entrenament per un nombre diferent d'arbres *ntrees* (valors de la forma $2^i,\ i = 1, ..., 10$) i ens quedem amb el que presenti un menor error de OOB (*Out Of Bag Error*). Els resultats obtinguts es troben a *Annex Taules 22, 23, 24, 25, 26, 27*.

5.2 Mesures de train

Taules 3, 4. Taules de l'error quadràtic mitjà i de la precisió d'entrenament. En el cas de la regressió s'expressa al valor del MSE i en els dos casos de classificació s'indica el percentatge d'encert.

Mat	GLM	LDA	Nv Bayes	KNN	ANNetwork	R. Forest
Reg.	3.554054	3.371622	11.10135	0	0.0202703	0.7482131
4 cat.	-	87.16216	82.77027	100	100	100
2 cat.	100	94.25676	88.17568	100	96.62162	100

Por	GLM	LDA	Nv Bayes	KNN	ANNetwork	R. Forest
Reg.	1.676955	1.259259	7.386831	0	0.0164609	0.4385165
4 cat.	-	85.80247	73.66255	100	100	100
2 cat.	92.38683	90.32922	86.41975	100	96.70782	100

S'observa com K-Nearest Neighbours és el que millor ha predit. No obstant, és un resultat que per ara encara no ens indica res. Això és degut a que hem realitzat el mètode amb $k = 1$ i l'error d'entrenament sempre serà nul. Per tant, no és un resultat significatiu.

Random Forest i Xarxes Neuronals semblen tenir una predicció de train excel·lent, amb un MSE molt petit. Caldrà veure si aquests resultats tant bons es mantenen a la validació.

5.3 Cross Validation

Taules 5,6. Taules de l'error quadràtic mitjà i de la precisió de validació (*cross-validation*).

Mat	GLM	LDA	Nv Bayes	KNN	ANN*	R. Forest
Reg.	3.99862	6.84241	13.45689	5.67540	3.1617	3.76446
4 cat.	-	71.27586	72.5977	71.90805	75.65433	79.91954
2 cat.	91.95402	85.63218	85.94253	88.31034	91.21125	92.98851

Por	GLM	LDA	Nv Bayes	KNN	ANN*	R. Forest
Reg.	1.74052	2.75604	8.32874	3.53958	2.72062	1.96378
4 cat.	-	79.56207	70.14881	71.78997	78.58468	81.41156
2 cat.	90.79082	84.65561	85.88435	83.03146	87.37968	90.99915

* Els valors de precisió de cross-validation de Xarxes Neuronals els hem obtingut amb el procediment de selecció dels valors dels hiper-paràmetres de l'apartat anterior, ja que era un cross-validation de 10x10.

Observem que els resultats empitjoren molt quan fem servir cross-validation. Atribuïm aquest fet a que el tamany de la base de dades d'entrenament a cada iteració de l'algorisme de cross-validation no és prou gran, ja que agafem nou dècimes parts de les dades d'entrenament, que ja eren ¾ parts del total de les dades. D'aquesta manera, ens quedem amb una base de dades reduïda amb la qual costa més extrapolar les solucions. De fet,

s'aprecia especialment si comparem l'error obtingut per regressió a les dues assignatures. Com havíem comentat, Portugués té gairebé el doble d'observacions que Matemàtiques. D'aquesta manera, observem que el MSE a tots els models utilitzats és molt inferior a Portugués. Considerem que es deu a que, al tenir més dades d'entrenament, s'obté una major extrapolació de les solucions i l'error disminueix. Al estar elevat al quadrat, la diferència entre l'error entre les dues assignatures s'accentua considerablement.

En general, s'aprecia un comportament molt bo per part de GLM i Random Forest.

6. Test i elecció del millor model

Taules 5,6. Taules de l'error quadràtic mitjà i de la precisió de test (amb les dades inicialment dividides en ¾ i ¼).

Mat	GLM	LDA	Nv Bayes	KNN	ANNetwork	R. Forest
Reg.	2.949324	3.969595	9.716216	1.584459	0.9898649	1.111042
4 cat.	-	82.43243	80.74324	90.54054	92.56757	95.94595
2 cat.	94.59459	90.54054	88.85135	95.94595	95.27027	97.97297

Por	GLM	LDA	Nv Bayes	KNN	ANNetwork	R. Forest
Reg.	1.401235	1.510288	7.518519	0.9465021	0.9588477	0.8475172
4 cat.	-	83.53909	72.63374	90.9465	91.97531	93.6214
2 cat.	92.38683	90.12346	87.2428	94.85597	95.06173	97.3251

En el cas de l'assignatura de matemàtiques, segons els resultats obtinguts, veiem que el model que millor s'ajusta a la regressió és el Xarxes Neuronals, ja que és el que té un menor Error Quadràtic Mitjà (MSE) amb un valor de 0,99. En canvi, per classificació en 2 i 4 categories el mètode que ens proporciona els millors resultats és el Random Forest, amb un 95,9% i 97.97% d'encert respectivament.

A l'assignatura de portuguès els resultats són molt similars. Observem que les millors prediccions s'obtenen amb el mètode de Random Forest. Per classificació, tenim un 93,62% i 97,33% de percentatge d'encert en els casos de classificació en 4 i 2 categories respectivament. Pel cas de regressió, un Error Quadràtic Mitjà (MSE) amb valor 0.8415638.

Per una banda, amb aquests resultats podem apreciar com els models tenen facilitat per predir en regressió a l'assignatura de portuguès. En canvi, veiem millors resultats en classificació en 4 categories en l'assignatura de matemàtiques, mentres que en classificació binària obtenim resultats molt similars. Això pot ser degut a que els alumnes de portuguès treuen notes en un rang menor als de matemàtiques que, com hem pogut veure als histogrames de l'apartat 3.2, estan més repartides.

Per altra banda, els resultats demostren que Random Forest és un mètode molt robust que funciona molt bé per extrapolar els resultats, és a dir, generalitzar solucions globals. Aquest aspecte l'observem en què la exactitud amb que predeix les notes d'ambdues assignatures per les dades de test és gairebé la mateixa que per les dades d'entrenament.

De la mateixa manera cal remarcar la qualitat de predicció que han demostrat els mètodes de KNN i Xarxes Neuronals, sobretot per regressió i per classificació binària. En canvi, Naive Bayes i LDA no han obtingut resultats tan molt satisfactoris en general.

Annex Imatges 21, 22, 23, 24, 25, 26.

Com podem observar a les imatges, segons l'algoritme de Random Forest, les variables explicatives que més importància tenen són G1 i G2. És molt coherent, ja que són les notes dels trimestres previs. És més interessant veure quines són les altres variables amb més importància.

Veiem que informació com el nombre d'assignatures suspeses anteriorment, les absències a l'escola o la intenció de realitzar estudis superiors prenen molta rellevància. També són importants l'edat i la salud de l'estudiant i els estudis i el treball de la mare.

7. Conclusions

I a la fi, considerem que hem complert tots els objectius del projecte i amb un gran èxit. Les nostres conclusions son de caràcter general i més aviat orientatives, ja que la base de dades ha estat petita i per tant les nostres conclusions estan subjectes a força incertesa. Per altra banda, el treball realitzat ha estat molt detallat i exhaustiu sobre la base de dades i les modalitats que aquesta ens ha permès explorar, com vostè pot comprovar en l'arxiu *Rmd* que se li adjunta en la carpeta del projecte.

Abans de prendre conclusions, volem deixar clar que aquest data set no ens ha semblat especialment objectiu ja que moltes de les variables que s'han utilitzat durant aquest estudi estan subjectes a l'opinió i sinceritat de l'alumne enquestat. Així doncs, creiem que amb dades més empíriques podrien trobar-se perfils molt més precisos que definíssin als alumnes. Variables com el propi consum d'alcohol, si no es mesuren en unitats objectives (per exemple, en nombre de llaunes o copes), és molt fàcil que els alumnes tinguin una concepció distorsionada sobre el seu consum. Altres variables com la relació familiar o bé la quantitat de temps lliure també són molt subjectives i depenen de la visió de l'estudiant. El fet de que existeixi aquesta subjectivitat fa que hi hagi molta incertesa. Per això, a l'hora de visualitzar les dades o intentar agrupar als estudiants pel seu rendiment acadèmic, la tasca és complexa i les conclusions força incertes.

Per una banda, a l'hora de realitzar l'anàlisi visual de les dades hem trobat relacions molt interessants. L'educació dels pares o la mida familiar no semblen influir en el rendiment acadèmic de l'estudiant, cosa que sí fa l'estat civil dels pares o el suport d'aquests en els estudis dels seus fills. També existeix una relació entre les qualificacions i la ocupació dels pares, així com haver acudit a una llar d'infants.
A més, hem observat que l'edat de l'estudiant és important respecte el rendiment acadèmic i, en canvi, el gènere o la salut no ho són. De la mateixa manera, s'associa viure a la ciutat, tenir accés a internet o un temps de desplaçament curt a uns bons resultats escolars. Si ens fixem en l'àmbit de lleure de l'estudiant, trobem que tenir relacions romàntiques, realitzar activitats extraescolars o consumir alcohol tenen molt d'impacte.
L'elecció de l'escola també és important, així com rebre ajuda educativa per part d'aquesta. Com era d'esperar, el temps dedicat a l'estudi és molt influent en el rendiment acadèmic, igual que ho són el nombre d'absències a classe o les assignatures suspeses anteriorment. En canvi, no sembla que les classes particulars tinguin un paper gaire rellevant a les notes. A més, i essent coherent, les qualificacions obtingudes als trimestres previs tenen una forta relació amb les notes finals.

La conclusió general que extraiem de l'anàlisis de les dades és que les notes obtingudes pels estudiants a l'assignatura de Matemàtiques estan molt més repartides que en les notes de Portugués (on sembla ser que els estudiants aproven amb més facilitat tal i com es mostra en els histogrames de la pàgina 5). Això podria semblar poc rellevant, però tal i com hem vist més tard a l'hora de efectuar prediccions ha tingut un impacte important i la classificació binària de l'assignatura de matemàtiques ha estat més senzilla que la de

Portugués, mentres que la regressió i la classificació en 4 grups ha estat més difícil i les prediccions lleugerament pitjors en l'assignatura de matemàtiques.

Per altra banda, pel que fa al modelat i predicció de dades, considerem que tot i la quantitat relativament petita d'observacions que teníem hem obtingut uns resultats molt més acurats del que es podria esperar. En la majoria de models (GLM, ANN, kNN o Random Forest) les prediccions han estat molt similars i l'error comès per aquests ha estat força petit. En els nostres experiments sembla ser que Random Forest ha estat el més consistent en pràcticament totes les situacions. No obstant, això no necessàriament vol dir que sigui millor mètode que la resta, si no que, per a les dades observades, aquest ha estat el model que millor ha funcionat.

En tot cas, el rendiment de GLM, del classificador de Naive Bayes, LDA, kNN i de les Xarxes Neuronals artificials ens ha semblat igualment excel·lent, i ens ha sorprès tenint en compte la quantitat tan reduïda de dades de la que disposàvem. Naturalment, Naive Bayes i LDA no han ofert bons resultats per regressió, tal i com era d'esperar, ja que no són mètodes pensats per fer classificació .

Per últim, hem de dir que la limitació fins a 14 pàgines imposada en l'enunciat ens ha privat de poder afegir molta feina que hem realitzat i que no hem pogut redactar amb tot el detall que ens hauria agradat.

Així doncs, aquí acaba el nostre projecte.

Agraïm al professorat de l'assignatura el temps que ens ha dedicat resolent els dubtes que ens han anat sorgint; hem gaudit enormement el procés de realització d'aquest estudi i de tot l'aprenentatge que hem adquirit al llarg del curs.

Patricia Cabot, Pau Lozano i Marina Rosell
Juny de 2020