**Predicció de notes Utilitzant Tècniques d’Aprenentatge Automatic**

Jan Planas Batllori

**Abstract—** En aquest article s’intentarà predir les notes del tercer trimestre dels alumnes de dues escoles portuguese. Hi ha dues assignatures sobre les que tenim informació sobre els alumnes i el temps i esforç que han gastat en cada una de les assignatures. El que s’intenta veure en aquest estudi és si hi ha possibilitat de poder predir les notes de forma bastant exacta o si afecta massa la capacitat de cada alumne a la hora d’avaluar-los. També intentarem veure si hi ha diferències entre les diferents assignatures, ja que hi ha matèries com les matemàtiques que sempre han sigut un mal de cap per a la majoria de gent durant la seva educació.

**Keywords—**School, math, Portuguese, students, alcohol cosumption, study, father, mother, grades, final grades.

─────── ♦ ───────

# 1 Introducció

Des de sempre la manera per avaluar els coneixements dels alumnes de ha sigut posant notes numèriques on, depenent del lloc, tenen un rang o un altre. Això sempre ha creat diferents devats ja que hi ha gent que creu que no és un mètode just degut a que no s’ajusta a la capacitat de cada persona. Mai us heu preguntat si de veritat el “factor humà” té importancia en aquest tipus de avaluacions? I si es així, es podrien predir les notes abans de fins i tot fer els examens? En aquest estudi es tractarà de respondre aquestes preguntes a més de intentar predir les notes de cada alumnes el més precís posible amb el mínim d’informació necessari,utilitzant models de regresió i basant-nos en mètriques com el r2\_score o el mean\_square\_error per a saber el rendiment del model. Les dades venen proporcionades per alumnes de dos col·legis portuguesos Mouzinho da Silveira i Gabriel Pereira on tenim diferents atributs com el temps d’estudi, la consumició d’alcohol, l’assistència a clase, les edats, … entre d’altres de dues assignatures diferents, matemàtiques i portugués.

# 2 Proposta/metodologia

Per a poder conseguir aquest repte es tractarà d’entrenar i avaluar diferents models de predicció de dades basant-nos sobretot el les mètriques r2\_score i mean\_aquare\_error dels diferents models. Però, abans de començar, haurem de fer una ullada a les dades que tenim sobre els alumnes per a veure si les dades ens serán útils i si están ben tractades. Primer es començarà per a fer una exploració de les dades i seguidament el tractament d’aquestes ja que per a entrenar els models necessitem tenir les dades processades i ben tractades.

Per a les prediccions es faran servir models de regressió ja que es tracta de predir un valor numèric. Com a models de predicció farem servir els següents: Regressió Lineal, Random Forest Regressor, Lasso, Riedge, Gradient Boosting Regressor i Support Vector Regression.

Per a l’anàlinis del rendiment de cada model es farà un K-fold cross validation per a veure el rendiment mitjà d’aquest i ens quedarem amb el que tingui millor r2\_score. Finalment farem un Grid Search per a buscar els millors hiperparàmetres del model i així aconseguir els millors resultats possibles.

## 2.1 Exploració de les dades

Primer de tot fem una primera exploració del data set i veiem que hi ha dos tipus de dades principalment: numèriques i categòriques. Fer aquesta distinció és important ja que cada tipus de dada es tracta de forma diferent. Podem veure que en les variables categòriques tenim paraules, les quals no són útils perquè no serveixen per entrenar els models així que haurem de fer un encoding per a que es puguin utilitzar. Per a millorar l’eficiència del model es normalitzarà les variables numèriques, és a dir, es ficaran totes en el mateix rang per a que el model pugui predir millor.

En una primera exploració es pot veure que en el data set no hi ha nans (missing values), i la majora de les dades segueix una distribució normal, que es lo esperat. En la figura 01 es pot veure la distribució de les notes del 1r, 2n i 3r trimestre.

Gráfico, Histograma

Descripción generada automáticamente

Fig. 01: Distribució de les notes dels 3 trimestres del curs.

Fixant-nos sobretot en les notes del 2n i 3r trimestre podem veure un augment en el número de 0, cosa que em sobta ja que pot suposar gent en casos especials que deixen l’escola o alguna altra cosa. Aquestes dades no afavoreixen per al model ja que fa que aquest li costi més trobar una relació bona entre les dades i aquests 0s es tractaran com outliers i es trauran del data set per a millorar en les prediccions.

## 2.2 Tractament de les dades

Com s’ha comentat es començarà per a fer un encoding de les variables categòriques, de les quals distingirem en dos tipus: binàries i no binàries. Per a les binàries es farà un LabelEncoder que assignarà 1 o 0 a l’atribut en funció de si és d’una classe o d’una altra. Aquest tipus d’encoding només es fa amb les binàries ja que si tinguéssim més classes podria afegir un ordre intrínsec que no existeix realment en les dades del dataset i es per això que pels atributs que no son binaris he decidit fer servir el OneHotEncoder. El OneHotEncoder crea columnes noves per a cada classe del atribut, el problema es que creix molt el tamany del data set i pot afectar després en el temps d’entrenament del model. Finalment es fa una normalització logarítmica de les dades numèriques i després s’escala amb el MnMaxScaler per a ficar les dades en un rang entre 0 i 1.

Per acabar es treuen els outliers comentats anteriorment i ja estem llestos per a poder entrenar els models i fer les nostres prediccions.

# 3 Experiments, resultats i anàlisi

Abans de començar a entrenar els models mirarem si els atributs tenen correlació amb el nostre target.

Escala de tiempo

Descripción generada automáticamente

## Fig. 02: Correlacions entre els atributs del datasets i els targets.

En la figura 02 es poden veure les correlacions de cada atribut amb els diferents targets (notes 1r, 2n i 3r trimestre) i es veu la importància de cada un en cada target. Es pot apreciar que els atributs que més correlació tenen son les notes entre sí, és a dir, que si treus bones notes al primer trimestre també les trauràs en el segon i segurament també en el tercer. Però abans de començar amb les prediccions finals intentarem veure si podem predir les notes del tercer trimestre sense cap de les notes dels trimestres anteriors.

## 3.1 Experiments

Texto

Descripción generada automáticamentePer a fer aquesta primera prova entrenarem una regresió lineal intentant predir els 3 targets diferents que tenim. També ho farem amb 3 datasets diferents: el de matemàtiques, el de portugués i un combinat dels dos. Això es fa per a com he mencionat al principi veure si pot existir algún tipus de dificultat en funció de la asignatura i veure si en funció dàquesta el model pot predir o no bé les notes.

Com es pot observar en la imatge els resultats obtinguts són pèssims amb una r2\_score máxima de 0.22, cosa que ens indica que el models no aconsegueix generalitzar bé.

Després d’aquest petit intent començarem a fer un seguit de proves per a veure el resultat final d’aquestes prediccions.

Primer de tot s’intentarà predir la nota del segon trimestre i la del tercer únicamente amb la nota del primer trimestre i després s’intentarà predir la nota del tercer trimestre amb les notes dels trimestres anteriors.

Texto

Descripción generada automáticamente

## Fig. 04: Resultats de predir notes del 2n i 3r trimestre en funció dels diferents datasets

En aquesta segona proba comencem a veure que els resultats d’aquesta són millors que els anteriors tenint com a r2\_score mínima un 0.64, cosa que es molt millor que el obtingut anteriorment. El millor resultat que obtenim es un r2\_score de 0.88 en el cas de predir les notes del tercer trimestre del data set de portuguès només. Es pot veure clarament on els resultats del data set de matemàtiques són més dolents que els altres, això em pot fer pensar que el fet de analitzar una assignatura o una altra afecta i que per predir el millor possible s’hauria de analitzar per separat. També em fa pensar en la possibilitat de que la dificultat de la assignatura afecta en les prediccions i la capacitat de cada alumne per cada una de les assignatures és un factor a tenir en compte a la hora de fer prediccions. El problema és que aquesta capacitat és difícil de mesurar i en el meu cas no tinc aquesta informació

Veient els resultats d’aquestes primeres proves finalmente es farà la predicció de la nota del tercer trimestre i s’utilitzarà les notes dels trimestres anteriors per aconseguir el millor resultat posible. Per a poder generalitzar bé s’utilitzarà els dos datasets combinats per dues raons: el tamany del dataset combina tés de 1044 i el del dataset de portugués de 600 només, perdriem la meitat dels valors i potser són pocs per a fer una bona predicció, l’altra raó es que volem generalitzar entre les dues assignatures i no centrar-nos en una de sola, ja que després es podría intentar fer servir per a altres assignatures.

Prier de tot es farà un K-fold cross validation amb els 6 models mencionats anteriormente i ens quedarem amb el que tingui millors resultats.

## 3.2 Resultats

Imagen de la pantalla de un celular de un mensaje en letras blancas

Descripción generada automáticamente con confianza baja

## Taula 1: Resultats del cross validation amb els 6 models diferents.

Es pot veure la taula 1 de resultats del k-fold cross validation amb els 6 models diferents. S’ha implementat un k-fold de 5 folds i s’ha fet la mitjana de cada resultat de cada fold i es mostra en forma de taula.

Veiem que hi ha 2 models que sembla que funcionen millor que els altres tenint unes r2\_scores de 0.837 i 0.838 que son de la regressió lineal i de Ridge respectivament.

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteFinalment ens quedarem amb el Ridge i posteriorment li farem un grid search per a trobar els millors hiperparàmetres i ajustar el model el màxim possible.

## Fig. 05: Diagrama de dispersió amb la línia de la regressió

En al figura 05 es mostra el gràfic de dispersió que mostra la relació entre els valors reals i els valors predits pel model. Podem observar que el model s’ajusta bastant bé però te complicacions a la hora de predir els 0. Això es pot causar a partir de la poca representació que hi ha de 0 en el data set.

# 4 Conclusions

Després de fer el grid search i tornar a entrenar el model aquest cop amb els nous hiperparàmetres obtenim els següents resultats.

Imatge que conté text, Font, nombre, captura de pantalla

Descripció generada automàticament

Fig. 06: Resultats del model Ridge amb hiperparàmetres

Podem veure que hem millorat bastant la r2\_score encara que també ha augmentat el mean\_square\_error . Això es pot deure a que el model generalitza millor (augment r2) però es veu més afectat pels valors extrems i com a conseqüència augmenta el mean\_square\_error.

Imatge que conté text, captura de pantalla, línia, Trama

Descripció generada automàticament

Fig. 07: Diagrama de dispersió amb la línia de la regressió

Es pot veure en el diagrama de dispersió de la figura 07 com tenim més valors extrems (al 0 o al 19-20) i això fa que augmenti el mean\_square\_error.

Com a conclusió podríem dir que a dia de avui tenim la capacitat de predir les notes dels alumnes en funció de atributs com les hores de estudi, la assistència, el col·legi al q vagin, ... però és essencial tenir un historial de les caràcterístiques del alumne més exhaustiu per a que aquestes prediccions siguin bones. Això ens deix clar que hi ha factors més importants a tenir en compte a la hora de fer aquestes prediccions com podrien ser la capacitat de memoritzar del alumne, les ganes i interès per la assignatura, la capacitat de realitzar exàmens, etc. A dia de avui es podria fer estudis més precisos i inclús es podria estudiar el cervell de cada estudiant ja que al final és el que més diferencia a les persones i les seves capacitats.

Finalment aclarir que aquest estudi serveix per a veure que les notes d’un estudiant depenen molt del caràcter i les capacitats d’aquest i això és únic de cadascú i només amb molt esforç i treball s’aconsegueixen millors resultats, encara que alguns tenen més facilitats que d’altres.

**Webgrafia**

[1] Scikit-learn. "Ridge Regression - scikit-learn 1.0.2 documentation." URL: <https://scikit-le-arn.org/stable/modules/generated/sklearn.linear_model.Ridge.html>

[2] Kaggle."Student Alcohol Consumption." URL: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

[3] Kaggle. "Best Notebook - 2.56 Mean Squared Error." URL: https://www.kaggle.com/code/deepak915/best-notebook-2-56-mean-squared-error

[4] OpenAI Chat. "OpenAI Chat - Conversational AI." URL: https://chat.openai.com