

Gerard Burgués Llavall

Joel Ferrando Ruiz

Predicció aplicant la regressió lineal

Anàlisi i predicció de les notes finals de l’alumnat

Table of Contents

[Anàlisi de dades 1](#_Toc86177502)

[Quines dades tenim? 1](#_Toc86177503)

[Quines dades farem servir? 4](#_Toc86177504)

[Objectiu 9](#_Toc86177505)

[Regressió 10](#_Toc86177506)

[Quins són els atributs més importants per una bona predicció? 10](#_Toc86177507)

[Que passa quan triem les nostres variables? 10](#_Toc86177508)

[Que passa quan triem les variables amb més correlació? 11](#_Toc86177509)

[Que passa quan dividim les dades en training i test? 11](#_Toc86177510)

[Quin conclusió podem extreure? 11](#_Toc86177511)

# Anàlisi de dades

## 

## Quines dades tenim?

La nostra base de dades tracta sobre el rendiment d’alumnes de secundària en dos escoles portugueses. Els atributs inclouen dades sobre les seves qualificacions, característiques demogràfiques, socials i característiques relacionades amb l’escola. Totes aquestes dades han sigut obtingudes de informes escolars i qüestionaris.

Anem a observar els camps dels que disposem

Pantalla de computadora

Descripción generada automáticamente con confianza baja

Volem saber quines dimensions té la nostra base de dades

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Podem observar que disposem de 395 files i 33 columnes.

Per obtenir més informació podem efectuar “describe() “on ens mostrarà informació com la mitja, el mínim, màxim, quartils ... En el nostre cas aquesta informació no ens serà útil ja que no la necessitem en la forma d’aplicar-ho però pots conèixer més informació sobre les dades on treballes.

Tabla

Descripción generada automáticamente

El següent pas a aplicar és saber si hi ha columnes amb molt valors NULL per poder eliminar-la ja que un gran percentatge de NULLS en una columna no ens serà útil.

També veurem quin tipus da dades es cada variable.

Tabla

Descripción generada automáticamente

Podem veure com no hi ha cap NULL en cap fila i el tipus de cada variable.

També podem saber quin valor poden agafar les variable de tipus “object” amb el següent codi.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

## Explicació sobre les variables

|  |  |
| --- | --- |
| **Variable** | **Explicació** |
| School | Escola en que es forma l’estudiant |
| Sex | Sexe de l’estudiant |
| Age | Edat de l’estudiant |
| Family Size | Mida de l’estudiant ( >3 o <3) |
| Parent Status | Situació parental (junts o separats) |
| Mother Education | Educació de la mare |
| Father Eductation | Educació del pare |
| Mother Job | Feina de la mare |
| Father Job | Feina del pare |
| Reason (to choose the school) | Raó per la qual triar l’escola |
| Guardian | Tutor de l’estudiant |
| Travel time (to school) | Temps que triga per anar a l’escola |
| Study Time | Temps que estudia |
| Failures (failed classes) | Classes que no ha anat |
| School Up (extra education) | Educació extra |
| Family Up (Family extra education support) | Educació extra familiar |
| Paid (Extra paid classes Math-Portuguese) | Ha pagat per extra classes |
| Activities (extra curricular ) | Activitats extracurriculars |
| Nursey (Attended to nursey school) | Ha atès a escola d’infermeria |
| Higher (wants higher education) | Vol educació superior |
| Internet (has internet) | Disposa d’internet |
| Romantic (in a relationship) | Forma part d’una relació romàntica |
| Family Relation | Calitat de la relació familiar |
| Free time | Temps lliure després de l’escola |
| Go out | Surt amb els amics |
| Day Alcohol Consumtion (during week) | Consumeix alcohol durant la setmana |
| Weekend Alcohol Consumption | Consumeix alcohol durant el cap de setmana |
| Health | Actual situació de salut |
| Absence | Número de cops que ha faltat a classe |
| G1 | Notes del primer període |
| G2 | Notes del segon període |
| G3 | Notes del tercer període |

## Quines dades farem servir?

Un cop observat i analitzat quina informació tenim hem de decidir quines dades ens seran útils i quines dades en podem prescindir.

A continuació es mostra aquesta taula:

|  |  |  |
| --- | --- | --- |
| **VARIABLE** | **ÚTIL** | **PER QUÈ?** |
| School | No | Les dades de les dos escoles no són representatives. Una escola es 90% mentre que l’altre només un 10%. |
| Sex | Si | El sexe es una bona mostra per predir si un noi/a tindrà més possibilitats de treure una nota més alta que un noi/a. |
| Age | Si | L’edat ens ajudarà a crear un model més concret |
| Family Size | No | Entenem que la mida de la família no influiria en el nostre model i per això la ignorem |
| Parent Status | No | Entenem que la situació parental no influiria en el nostre model i per això la ignorem |
| Mother Education | No | Entenem que la educació de la mare no té per que influir en un estudiant. La ignorem |
| Father Eductation | No | Entenem que la educació del pare no té per que influir en un estudiant. La ignorem |
| Mother Job | No | Entenem que la feina de mare no té per que influir en un estudiant. La ignorem |
| Father Job | No | Entenem que la feina de pare no té per que influir en un estudiant. La ignorem |
| Reason (to choose the school) | No | La raó per que triar una escola no ens dona cap informació útil sobre la futura nota de l’alumne |
| Guardian | No | El tutor no ens influeix en una futura predicció. Optem per ignorarlo |
| Travel time (to school) | No | El temps d’un estudiant cap a l’escola entenem que no dona cap motiu per aprovar o suspendre, l’ignorem. |
| Study Time | Si | El temps que un estudiant dedica en estudiar es necessari per la futura predicció de les notes |
| Failures (failed classes) | Si | Les classes que ha suspès potser un avanç de les notes que pot treure un estudiant |
| School Up (extra education) | Si | Un estudiant amb més recursos pot influir amb les futures notes de l’estudiant, per tant, necessari. |
| Family Up (Family extra education support) | No | Educació familiar extra creiem que es necessari ja que pot afectar a les futures notes. |
| Paid (Extra paid classes Math-Portuguese) | Si | Extra classes poden beneficiar a l’estudiant a treure millors resultats |
| Activities (extra curricular ) | Si | Com més activitats tingui un estudiant pot afegir o treure facilitats, per tant, ho tenim en compte |
| Nursey (Attended to nursey school) | No | Atendre a classe d’infermeria no afecta a les futures notes . |
| Higher (wants higher education) | Si | Voler o no extra educació pot influir en la decisió d’estudiar més o menys, per tant ho tindrem en compte |
| Internet (has internet) | No | Tenir internet no hauria d’influir en la quantita d’hores que s’estudia. No ho considerem necessari |
| Romantic (in a relationship) | Si | Estar en una relació no ens ajudarà a predir si un estudiant treu millors notes |
| Family Relation | Si | La relació amb la família pot afectar als resultats de l’estudiant. |
| Free time | Si | Més temps lliure pot provocar relax o més estudi. Afecta al resultat final. |
| Go out | No | Considerem que sortir amb els amics no seria necessari tenir-ho en compte |
| Day Alcohol Consumtion (during week) | No | Les dades oferides en aquesta variable son desproporcionades |
| Weekend Alcohol Consumption | Si | Creiem que l’accés d’acohol pot afectar al resultat de l’estudiant |
| Health | Si | Creiem que la situació de salut pot afectar a les notes de l’estudiant. |
| Absence | Si | Un estudiant que falti varios cops a classe ho pot veure afectat en la nota, per això ho tenim en compte |
| G1 | Si | Les notes dels períodes anteriors són necessaris per predir un futur |
| G2 | Si | Les notes dels períodes anteriors són necessaris per predir un futur |
| G3 | Si | Les notes dels períodes anteriors són necessaris per predir un futur |

Una vegada ja hem decidit quines variables utilitzarem i quines no, hem d’eliminar-les

data = data.drop(['school', 'famsize', 'Pstatus', 'Fedu', 'Medu', 'Fjob', 'Mjob', 'reason', 'guardian', 'traveltime', 'famsup', 'nursery',  'internet', 'goout', 'Dalc'], axis = 1)

Mirem a veure si s’han eliminat

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto, Sitio web

Descripción generada automáticamente

Ja estan eliminades de la nostra base de dades.

Volem conèixer més encara les dades, mirarem la correlació entre els atributs d’entrada.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Aquesta correlació que es mostrarà més endavant amb més detall ens ajudarà a veure i a decidir quines dades hauríem d’utilitzar per la regressió.

## Distribució

Volem saber quina distribució segueixen algunes de les nostres variables. Per això podem mostrar la següent gràfica en el notebook:

Gráfico, Calendario

Descripción generada automáticamente

En aquest exemple només mostrem una part, però podem observar totes les variables amb les quals treballem i les seves relacions.

Analitzant-les podem veure que :

* G2, free time i goout tenen una distribució gaussiana

Chart, histogram

Description automatically generated

* Failures, trabeltime, dalc, abscence tenen una distribució geomètrica

Histogram

Description automatically generated

## Transformar variables

Per poder aplicar el model correctament necessitem que les variables amb “String” tinguin un número associat. Procedim a canviar-les de la següent manera.

Texto

Descripción generada automáticamente

# Objectiu

El nostre objectiu serà fer una predicció utilitzant el mètode de Descens de Gradient per predir la nota que un estudiant tindrà l’últim curs. Utilitzarem totes les dades anteriors (“Si”), tractarem les caselles buides si n’hi ha, normalitzarem les dades i mostrarem un anàlisi gràfic de les diferents variables.

# Regressió

## Quins són els atributs més importants per una bona predicció?

Els atributs més importants són aquells que tenen una correlació més gran amb la variable objectiu. Tenir una correlació gran indicarà que la variable objectiu es veu greument influenciada per la variable donada.

## Que passa quan triem les nostres variables?

Com veiem a la següent imatge que representa la correlació de les variables amb la nostra variable objectiu, només es poden observar 2 variables amb correlació més alta de 0.8 respecte la ‘G3’.

Les variables amb més correlació són G1 i G2 que representen les notes dels cursos anteriors. Aquestes variables seran essencials pel càlcul del nostre regressor.

![Chart, timeline

Description automatically generated]()

Chart, scatter chart

Description automatically generatedEs pot veure en el gràfic entre la predicció i els valors de G2 que la predicció es totalment errònia. Això es a causa de les variables que no tenen una correlació. Ens alteren els resultats d’una forma molt elevada.

**Mean Squared Error** : 0.00837

**R2 Score**: 0.8399

## Que passa quan triem les variables amb més correlació?

Utilitzant les variables amb més correlació, es a dir ‘G1’ i ‘G2’ ens dona com a resultat aquests valors:

Chart, scatter chart

Description automatically generated

**Mean Squared Error** : 0.0093

**R2 Score** : 0.0822

### Que passa quan dividim les dades en training i test?

Quan seleccionem un training de 0.85 ens diu que l’error en l’atribut G1 te un **error MSE de 0.02** mentre que el MSE de **l’atribut 2 és de 0.008**. Com s’ha pogut veure a la gràfica “heatmap” podem afirmar que com més correlació la variable té amb l’objectiu menys error s’obté.

## Quin conclusió podem extreure?

Com podem observar el MSE\_15 Variables és més petit que el MSE\_2 variables. Això pot passar per dos raons:

1. La informació que ens proporcionen les dades de les 15 variables ens estan provocant soroll i donant informació a la regressió que no és rellevant. Això pot provocar que hi hagi una millora però no és significativa.
2. La informació proporcionada ens aporta un grau molt petit d’informació que la linear regression troba significativa.

No hem pogut treure quina es la variable que té més informació significativa per això nosaltres ens decantem a que les dades ens estan aportant soroll i per tant irregularitats al model.