Pràctica 1. Regressió.

<https://www.kaggle.com/datasets/dipam7/student-grade-prediction>

Pol Reyes

Serena Sánchez

Valentí Torrents

INDEX

1. Introducció
   1. Explicació base de dades
   2. Motivació i objectius
2. Apartat C
   1. Dades
   2. Distribució Gaussiana
3. Apartat B
4. Apartat A

A LA WEB POSA QUE ES MATES + PORTUGUES PERO NOMÉS HI HA LES NOTES DE MATEMATIQUES PER DECARREGAR-SE TOT I DIR QUE HI HA 2 DATASETS.

**Introducció**

La base de dades que es treballa en aquest informe esta proporcionada per Paulo Cortez, de University of Minho, GuimarÃ£es dins de UCL Machine Learning Repository.

Les dades fan referència al rendiment dels estudiants de dues escoles: Gabriel Pereira i Mousinho da Silveira. Més concretament tracta sobre les notes finals de matemàtiques dels tres períodes que comprenen un curs en l’educació secundària a Portugal on G1 i G2 són les notes referents a cadascun dels períodes i G3 és la nota final.

Cada mostra fa referència a un alumne d’alguna d’aquestes escoles on tenim diferents atributs amb informació sobre: on viuen, quina edat tenen, dades relacionades amb el seu cercle familiar, la seva situació socioeconòmica i també sobre els seus resultats acadèmics.

Les estadístiques coloquen a Portugal a la cua d’Europa quan parlem de tasa de fracàs estudiantil, en concret, en referència a les assignatures de Matemàtiques i Llengua Portuguesa.

La motivació, per tant, d’aquest estudi es intentar saber quins atributs són més rellevants a l’hora de predir la nota final d’un curs, per tal de poder centrar en aquests més atenció i recursos, i així millorar la qualitat de l’educació. Per tant, l’experiment vol respondre, bàsicament preguntes com: ¿Quins factors ajuden a predir el rendiment escolar? ¿Quins atributs no afecten en aquest? ¿És possible predir el rendiment dels estudiants?.

A la base de dades que se’ns proporciona hi ha tres atributs que poden ser objecte d’estudi i per als quals volem fer la predicció: G1, G2, G3 explicats prèviament. Considerem que l’atribut objectiu ha de ser G3, la nota final del curs.

De la mateixa forma, hem decidit utilitzar G1 i G2 com a dos característiques més de la base de dades, ja que creiem que aquests dos atributs poden ser determinants per predir la nota de G3. Sobre aquesta hipòtesi, farem les primeres proves de predicció de G3.

Per una altra banda, en una segona part de l’informe es tractarà d’intentar predir G3 sense tenir en compte G1 i G2. La idea surt del següent raonament: l’objectiu és utilitzar les dades per saber quins recursos podem utilitzar per millorar el rendiment acadèmic. Si basem la predicció de la nota final en dues notes prèvies a aquesta, no tindrem les eines per complir els objectius, ja que la conclusió serà que s’han de treure millors notes als altres trimestres. Traient G1 i G2, es deixa que els altres atributs tinguin més importància al model final i que es pugui treballar de forma més acurada els objectius.

**APARTAT C**

**Dades**

A continuació mostrarem un llistat amb tots els atributs acompanyats del seu tipus:

| **Atribut** | **Descripció** | **Domini** |
| --- | --- | --- |
| **sex** | sexe del estudiant | [masculi, femeni] |
| **age** | edat de l’estudiant | de 15 a 22 |
| **school** | escola de l’estudiant | [Gabriel Pereira o Mousinho da Silveira] |
| **address** | zona on viu l’estudiant | [urbana, rural] |
| **Pstatus** | cohabitatge dels pares | [junts, separats] |
| **Medu** | educació de la mare | de 0 a 4 |
| **Mjob** | feina del la mare | [professora, àmbit de la salut, serveis, a casa, altres] |
| **Fedu** | educació del pare | de 0 a 4 |
| **Fjob** | feina de la mare | [professora, àmbit de la salut, serveis, a casa, altres] |
| **guardian** | tutor legal | [mare, pare, altres] |
| **famsize** | tamany de la familia | [0 si <=3,1 si > 3] |
| **famrel** | qualitat de les relacions en familia | del 1 - molt dolenta al 5 - molt bona |
| **reason** | per que es va escollir l’escola | [proximitat, reputació, preferencia de curs, altres] |
| **traveltime** | temps del trajecte a l’escola | de 1 - 15 minuts al 4 - més d’una hora |
| **studytime** | hores d’estudi per setmana | de 1 - < 2 hores al 4 - més de 10 hores |
| **failures** | número d’assignatures suspeses al passat | de 1 al 4 - 4 o més |
| **schoolsup** | classes de suport extres | [si, no] |
| **famsup** | suport familiar en l’educació | [si, no] |
| **activities** | realització d’activitats extraescolars | [si, no] |
| **paid** | classes extres pagades | [si, no] |
| **internet** | accés a internet a casa | [si, no] |
| **nursery** | atés a l’enfermeria de l’escola | [si, no] |
| **higher** | vol cursar estudis superiors | [si, no] |
| **romantic** | esta a una relació amorosa | [si,no] |
| **freetime** | temps lliure despres de l’escola | de 1 - molt poc al 5 - molt |
| **goout** | sortides amb els amics | de 1 - molt poc al 5 - molt |
| **Walc** | consum d’alcohol els caps de setmana | de 1 - molt poc al 5 - molt |
| **Dalc** | consum d’alcohol al dia a dia | de 1 - molt poc al 5 - molt |
| **health** | estat de salud | de 1 - molt poc al 5 - molt |
| **absences** | número de faltes a l’escola | de 0 a 93 |
| **G1** | nota del primer període | de 0 a 20 |
| **G2** | nota del segon període | de 0 a 20 |
| **G1** | nota final | de 0 a 20 |

*Taula 1. Atributs de la base de dades.*

**Anàlisi de les dades**

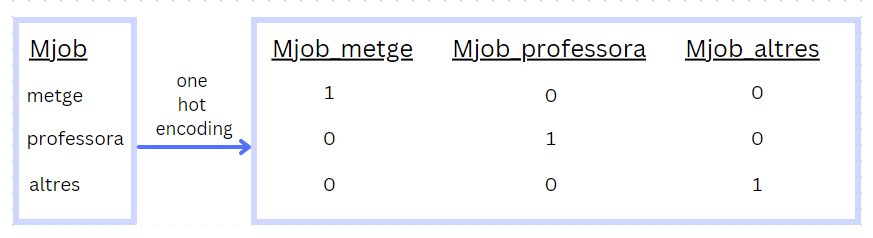
El primer problema que ens trobem quan volem començar a treballar amb aquesta base de dades és la quantitat d’atributs amb valor categòric, per exemple: la feina del pare o la mare, o el número de persones al nucli familiar. Aquests tipus de variables no ens permeten dur a terme operacions numèriques ni entrenaments de regressió. Per tant, abans de posar-se a estudiar les dades, s’han hagut de modificar algunes columnes de la base de dades.

Per una banda, s’han convertit a enters aquells atributs amb domini categòric binari de la que veiem a la següent taula.

| school | sex | address | famsize | Pstatus | Atributs amb domini [‘yes’, ‘no’] | Canvi a int |
| --- | --- | --- | --- | --- | --- | --- |
| GP | F | U | GT3 | A | no | 0 |
| MS | M | R | LE3 | T | yes | 1 |

*Taula 2. Modificació a enter del domini dels atributs categòrics.*

Per altra banda, s’ha creat una funció que tracta les dades de tipus categòriques amb més de dues opcions al domini per poderles utilitzar als algorismes. Això s’ha dut a terme mitjançant el mètode One Hot Encoding als atributs Mjob, Fjob i guardian. Aquesta estratègia consisteix en crear una columna per cada resposta diferent que aparegui en un atribut i, per cada registre, posar un 1 a la columna a la que pertany aquest registre i la resta posar 0’s.



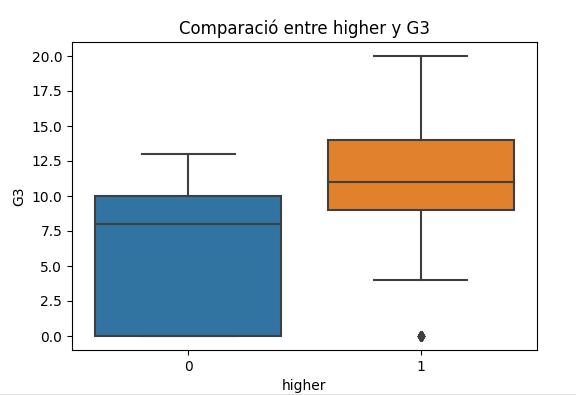
*Imatge 1. Exemple de one hot encoding aplicat a l’atribut Mjob*

**Eliminació d’atributs amb poca informació**

Primer de tot volem assegurar-nos que les dades que tenim a la base de dades ens aporten algun tipus d’informació. Per dur a terme això, analitzarem els atributs que considerem més o menys importants, la distribució d’alguns d’ells i si dels que són de tipus numèric n’hi ha que tinguin molts valors nuls o 0.

Abans de posar-se a manipular les dades s’ha decidit eliminar els següents atributs: *school*, *address*, *reason*, *nursery*, ja que no pensem que cap d’ells tingui, molta relació amb l’objecte d’estudi.

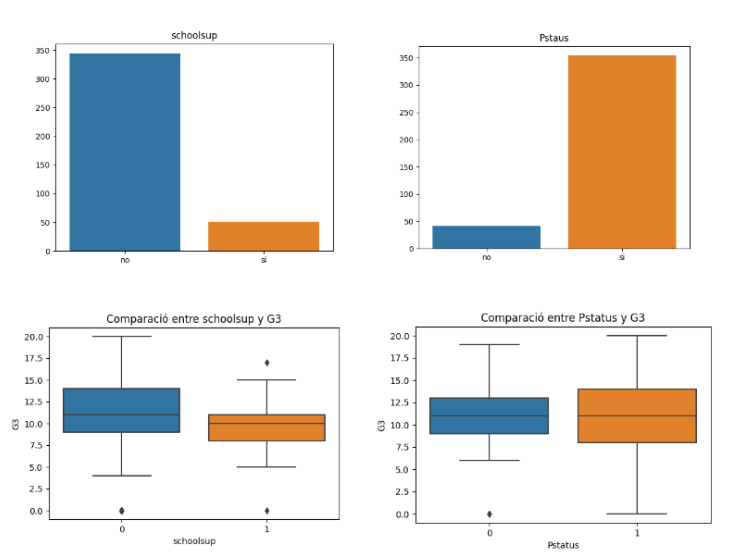
Per les variables amb domini binari, farem un seguit de gràfiques per saber si podem seguir descartant alguns atributs de la base de dades. Per exemple, atributs que no ens aporten informació són aquells en que la majoria de les mostres són 0 o 1.

L’atribut on major diferència hi ha entre les mostres que són ‘yes’ i les que son ‘no’, és l’atribut higher on només el 5% de les mostres diu que ‘no’. Pensem que, tot i haver poques mostres amb l’atribut higher a 0, és un atribut important per determinar l’esforç acadèmic. Per tant, abans d’eliminar aquest atribut realitzarem un diagrama de caixes per veure la distribució respecte la nota G3. 

D’aquesta forma, veiem que aquest atribut si influeix a l’hora de predir la nota de G3, si més no, a l’hora de determinar si s’aprova o es suspèn l’assignatura de matemàtiques. Per tant, no eliminarem aquest atribut de la nostra base de dades, al contrari, aplicarem el mètode oversampling per tenir el triple de mostres amb aquesta característica.

*Gràfica 1. Distribució de l’atribut higher en funció de G3*

Els altres atributs on hi ha molta diferència entre els 0’s i els 1’s es a Internet, schoolsup, famsize i Pstatus, per tant, fem el mateix procediment que amb higher. A les gràfiques de sota en posem dos exemples.

*Imatge 2. Exemples de desbalancejament de dades i distribució.*

S’ha decidit no descartar ni schoolsup ni internet pel mateix motiu que higher, ja que la distribució en funció de G3 no es tan semblant com per pensar que aquest atribut no ens aportarà informació. Però sí que es descarten els atributs Pstatus i famsize, ja que la nota final no té gaire relació amb aquests.

Després d’aquest anàlisi, transformació i modificació de les dades del nostre estudi, obtenim una base de dades de 36 atributs.

**Distribució Gaussiana**

En l’estudi d’una regressió lineal, que és el model final que volem aconseguir per complir amb els objectiu establerts, és important tenir dades que segueixin una distribució normal. Això és degut a que les dades s’escalaran en un interval petit en concret i farà que tots els atributs, en un principi, tinguin un pes similar al model. Per tant, permet comparar i evaluar dades i no treballar amb magnituds molt diferents que desbalancejarien les dades i dificultarien el seu anàlisi.

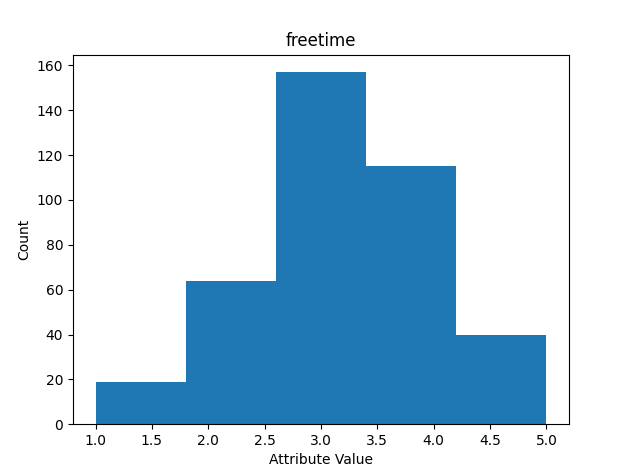
Degut a això, ens interessa esbrinar quins dels atributs segueixen una distribució Gaussiana; primer de tot s'ha descartat aquells atributs amb domini binari ja que aquests no poden seguir una distribució normal.

Dins de la pàgina web de les dades, trobem un histograma per cada atribut:



*Imatge 3. Histogrames de la pàgina web de la base de dades.*

Dels casos que sembla que segueixen una distribució normal (freetime, goout, G1, G2), s’ha realitzat un estudi de distribució normal amb la funció de normalització normaltest de Python, per confirmar que aquests atributs segueixen aquesta distribució. Finalment, l’únic atribut que ha resultat seguir una forma normal amb una alpha de 0.05 es l’atribut de freetime.



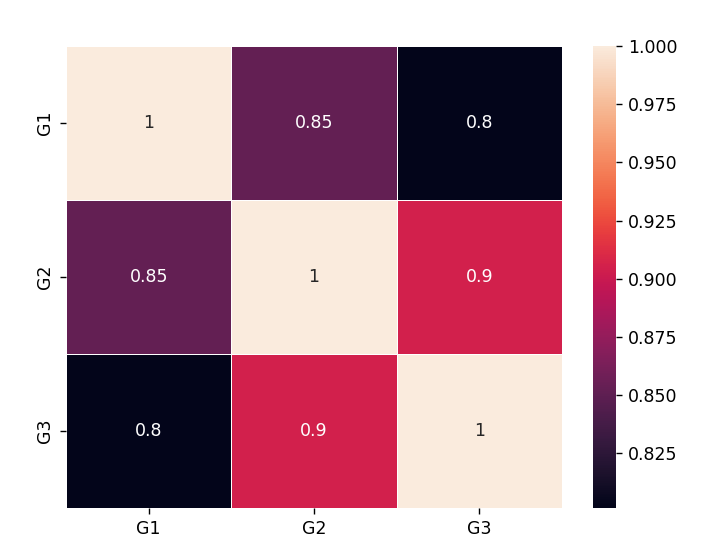
*Gràfica 2. Histograma de l’atribut freetime.*

Amb això podem avançar-nos als possibles resultats de l’informe i, tenir en compte, que les dades proporcionades de la forma en que les tenim, potser no segueixen una distribució adequada per l’estudi d’un regressor. Tot i així, els atributs G1, G2 i G3, segueixen una distribució bastant semblant a la normal, fet que ens facilitarà crear models amb aquestes dades.

**Correlació entre els atributs**

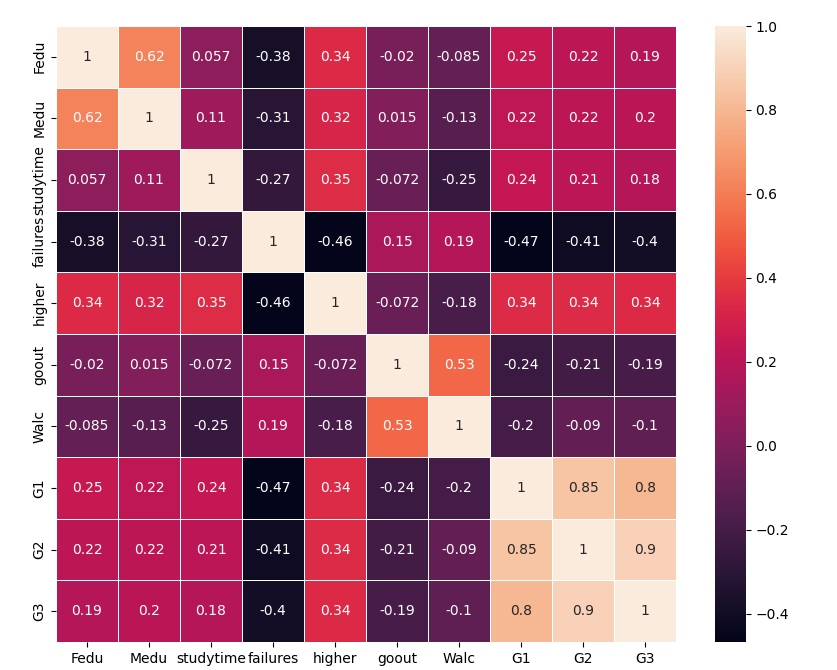
Per conèixer de forma més precisa la relació entre els diferents atributs de la base de dades s’han utilitzat dos mètodes de visualització diferents: la matriu de correlació i unes gràfiques de relació.

Matriu de correlació



Envers les matrius de correlació, s’ha realitzat la visualització dels atributs per grups. La correlació més alta s’ha trobat en els atributs ‘G1’, ‘G2’ i ‘G3’ com podem veure a la imatge de la dreta. Aquesta relació es interessant ja que el nostre atribut objectiu es ‘G3’, per tant, això ens pot donar indicacions sobre que informació necessitem per predir de manera correcta la nota final.

Tenint en compte això, s’ha fet la visualització per grups dels atributs, però sempre tenint en compte el conjunt d’atributs ‘G1’, ‘G2’ i ‘G3’, d’aquesta forma, podem veure si algun altre atribut es pot relacionar amb ‘G3’ i, si no és el cas, podem veure els que es relacionen amb ‘G1’ i ‘G2’, de forma que fent combinacions d’aquests atributs que en principi no tenen molt a veure amb l’atribut objectiu, podem arribar a una relació amb aquest, relacionant-los de forma transitiva.



En general, deixant de banda ‘G1’ i ‘G2’, els factors de correlació amb G3 són bastant baixos, això pot dificultar el trobar un model bo per al nostre estudi que s’adapti als nostres objectius.

Tot i així, fem una visualització dels atributs que, després de ‘G1’ i ‘G2’ están més correlacionats amb ‘G3’:

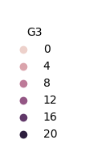
Com la majoria de dades que tenim son binaries i no poden seguir una distribució normal i, conseqüentment tampoc son les dades més adequades per realitzar el nostre model de regressió hem decidit reduir el dataset al número d’atributs que veiem a la imatge de la matriu de correlació.

Canvi de dataset

En un primer estudi del dataset, es va intentar aconseguir un model que tingués els 36 atributs de la base de dades i els resultats no van ser els esperats. Al realitzar el regressor multivariat, els coeficients d’alguns atributs es disparaven molt envers els altres, no hi havia una bona representació amb PCA del model i, per tant, els resultats no resultaven ser els esperats i no ens eren útils.

En un segon estudi del dataset, s’ha provat ser més estrictes amb els atributs que utilitzem i s’ha decidit crear un nou dataset de 9 atributs amb aquells amb els que la correlació amb G3 és aproximadament 0.2 o més (el notebook està actualitzat amb aquest segon dataset -df\_ca-). Durant l’informe, s’aprofundirà més en aquest segon, però en el següent apartat es faran comparacions dels resultats d’ambdòs, per veure si hem aconseguit millores amb aquest canvi.

Pairplot

D’una forma semblant a les matrius de correlació i tenint en compte els resultats d’aquestes, s’ha realitzat la visualització de distribucions amb la funció pairplot per la combinació dels atributs més relacionats en funció del valor de ‘G3’.

*Gràfica 3. Pairplot de les distribucions de la relació entre els diferents atributs.*

A la imatge podem veure la distribució que segueixen tots els atributs. Si ens fixem a les 4 gràfiques de la cantonada dreta abaix, trobem la distribució de G1 i G2, entre ells i amb G3. Fàcilment es pot intuir una recta que podria donar bons resultats com a model de predicció.

**APARTAT B**

**Primeres regressions**

Regressió lineal simple

En una regressió lineal, l’objectiu és trobar un model, és a dir, una recta on podem trobar una relació precisa entre la predicció que fem i la resposta real.

Per tant, si adaptem l’equació de la recta a la del nostre model ens queda el següent:

On y es la variable dependent, és a dir, el valor a predir i aquest depèn linealment de la variable independent x1. Les a’s es denominen coeficients del model o paràmetres, són aquests els valors que variaran per tal de trobar la millor predicció possible.

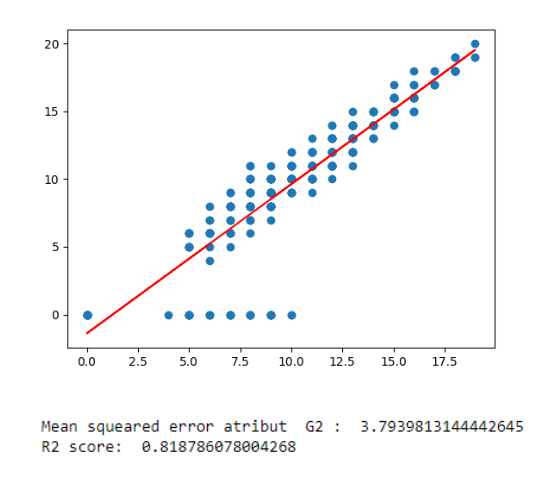
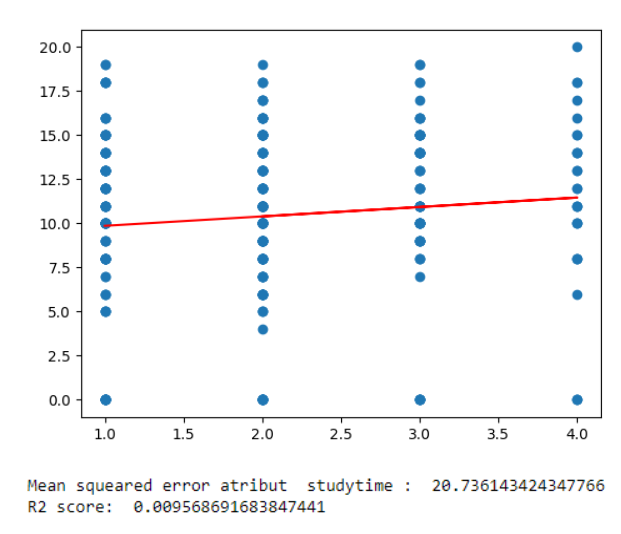
En el nostre cas, un dels models lineals que podem construir respondria a la següent pregunta: depèn la nota del final de l'alumne en matemàtiques de les hores que estudia a casa per setmana? La nota final de l’alumne seria la nostra variable dependent (y) i x1 les hores que hi dedica a casa. D’aquesta forma, es podrà crear un model que ens permeti predir si algú treura una nota alta o baixa segons les hores que estudia a casa.

Es poden crear molts models diferents, però l’objectiu és saber quin és el millor model que podem escollir per predir aquesta resposta que busquem.

Per tant, aquest apartat consisteix en aconseguir aquest millor model. Per aconseguir això s’han utilitzat dos evaluadors diferents: l’error quadràtic mitjà i R2-score. La primera mètrica consisteix en determinar la diferència entre el que prediu el nostre model i el valor real, s’eleva aquesta diferència al quadrat i es fa la mitja. Ens interessa aconseguir un valor proper al 0, tenint en compte que el resultat sempre serà positiu. En canvi, la mètrica R2-score també tracta de mesurar la distància des dels punts del groundtruth fins la recta determinada pel model per mitjà de la variança, l’objectiu és que doni un valor proper a 1. Tot i així, el valor d’aquesta mesura no indica si el model de regressió és adequat o no.

Regressió lineal simple sense normalització de dades

Probem amb els 9 atributs el regressor lineal i obtenim el millor resultat amb l’atribut ‘G2’, on veiem una recta vermella que d’adapta bastant bé al model i el pitjor amb l’atribut studytime on tenim un atribut binari que no podem adaptar al regressor de forma individual.



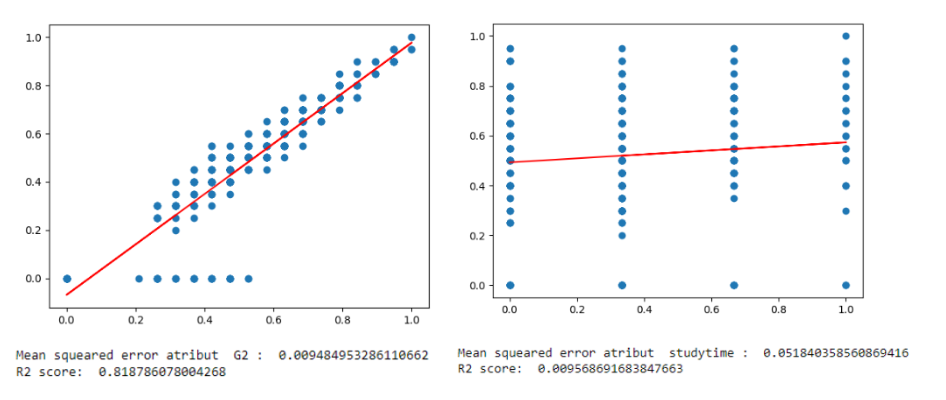
Procediment de normalització de les dades

En aquesta base de dades tenim atributs en rangs diferents, és a dir, tenim molts atributs amb domini binari, d’altres que prenen valors entre el 0 i el 5, altres, com les notes, que prenen valors del 0 al 20 i l’atribut absences que té un rang més ampli pot arribar fins a 90. D’aquesta forma, operar amb aquest diferència de dominis pot afectar als resultats del regressor i donar resultats menys bons que si tinguéssim un rang de dades similars. Això és degut, a que el regressor li donarà més importància a l’atribut amb valors més grans, llavors el càlcul del MSE perdrà sentit.

Llavors, per tal d’escalar les dades en un rang més petit i similar i poder aplicar el regressor per obtenir bons resultats, s’ha fet un procés de normalització amb la funció preprocessing.MinMaxScaler() de totes les dades. Ara, tornem a provar el regressor.

Regressió lineal simple amb normalització de dades

Tot i que amb mse i r2-score escalats a la normalització de les dades, obtenim els mateixos resultats.



Regressió lineal multivariada

En aquest apartat, s'estén la regressió a un polinomi amb més d’una característica. En el nostre cas, el model inicial tindrà 9 característiques (x0,...,x9).

Per a realitzar aquest model, primerament, utilitzarem el primer dataset (35 atributs) i el segon dataset (9 atributs). Després, partint del segon dataset, anirem entrenant el model sense aquells atributs on el pes que l’acompanya, un cop ha convergit l’algorisme, sigui molt petit.

Després, agafarem els 4 atributs més correlacionats amb G3 (ho veiem a la matriu de correlació) i entrenarem amb aquests el regressor. De totes aquestes probes en recollirem el MSE, obtenint la següent taula:

| dades | no normalitzades | normalitzades |
| --- | --- | --- |
| primer dataset | 3.268 | 0.008 |
| segon dataset (9 atributs) | 3.669 | 0.009 |
| segon dataset (6 atributs) | 3.706 | 0.009 |
| segon dataset (4 atributs) | 3.707 | 0.009 |
| atribut G2 | 3.793 | 0.009 |

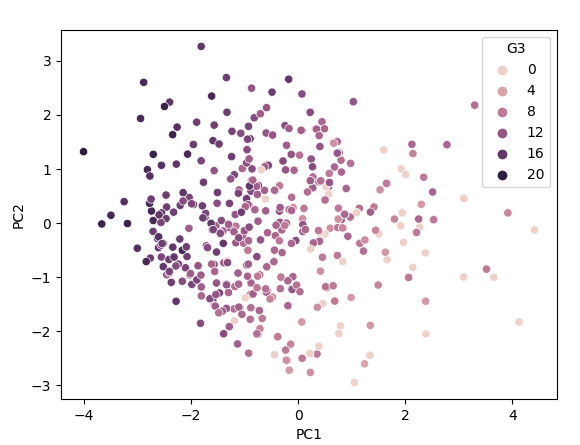
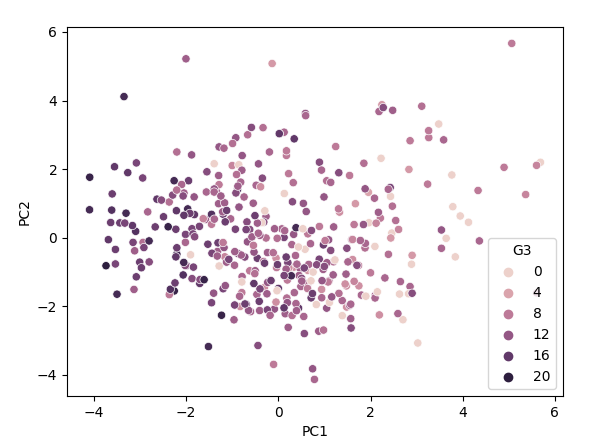
Com podem observar, els millors resultats els obtenim amb les dades normalitzades. No hi ha gaire diferència entre l’ús d’un dataset o altres. Al següent apartat intentarem visualitzar el model del primer i del segon dataset mitjançant una anàlisi de components principals.

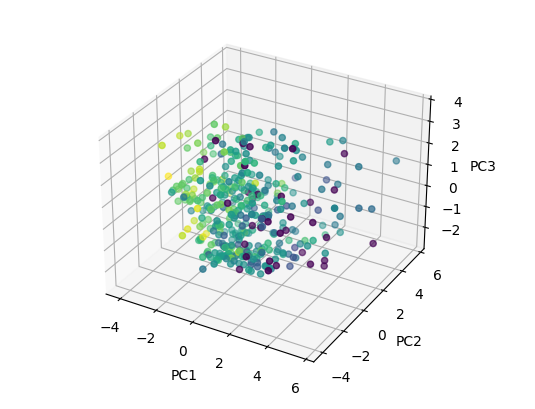
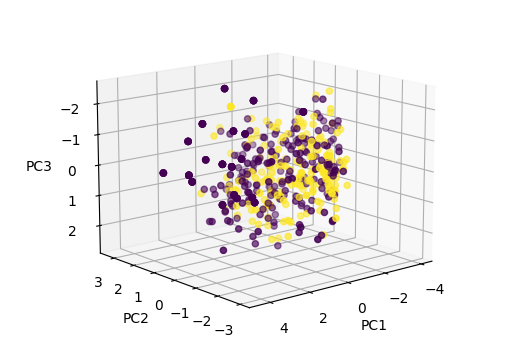
Anàlisi de components principals

Aquest és un mètode d’anàlisi de dades multivariades que permet estudiar i visualitzar conjunts de dades multidimensionals. Consisteix en projectar mostres d’un espai p-dimensional amb número de variables p en un espai k-dimensional on k < p de forma que es conservi el màxim d’informació possible. D’aquesta forma, quan es treballa amb un conjunt de dades grans i volem contextualitzar-les en un espai (com pot ser una gràfica), es redueix la dimensionalitat aplicant un PCA que es queda amb 2 o 3 components principals de variables no correlacionades permetent així la visualització de les dades.

Pel nostre conjunt de dades, actualment tenim 36 atributs, per tant, la visualització d’un model que tingui en compte tants atributs no és possible. Davant d’aquesta situació, primerament s’aplicarà un PCA amb dimensió 2, que reduirà el número d’atributs a 2 components principals i s’aplicarà un regressor logístic per tal de saber si aquestes components ens permeten construir un model que predigui la nota G3 amb el mínim d’error possible.

S’ha estudiat també la variança i, tot i seguir sent baixa, en el model de l’esquerra (primer dataset) la variança d’aquestes dues components principals envers a totes les altres era de 0,1. En canvi, al model de la esquerra (segon dataset), la variança es d’un 0.35. De fet, si observem aquesta imatge, els punts del mateix color es distribueixen en gran part en una zona concreta.



Repetim el procés però amb dimensió 3. La visualització obtinguda resulta difícil d’interpretar i no ens aporta gaire informació. Això pot tenir a veure amb el tipus de dades que estem treballant i que la correlació entre aquestes sigui baixa. 

Eliminem G1 i G2

Com a prova extra al model de regressió lineal multivariada, entrenem l’algorisme amb els atributs més correlacionats, però, eliminant els atributs G1 i G2 que fins ara han tingut els millors resultats en totes les proves.

Amb aquest model, aconseguim un MSE de 0.043, bastant més gran que l’error que aconseguim tenint en compte G1 i G2. Els atributs que passen a tenir més pes quan entrenem aquest model son: failures i Medu.

**APARTAT A**

**Descens del gradient**

Com s’ha comentat anteriorment, per tal d’establir un model que predigui de forma correcta l’atribut objectiu, es necessita una recta amb uns paràmetres en concret. El procés automatitzat d’escollir els paràmetres més convenients pel model s’anomena: Descens del Gradient.

De forma ràpida, aquest algorisme inicialitza els paràmetres de la funció de regressió i s’estableix una funció de cost que calcula la suma dels errors entre el conjunt d’aprenentatge i la funció de regressió. Llavors, es van fent modificacions dels paràmetres dins que la funció de cost (J) s’estabilitzi en el valor més petit possible, tenint com a objectiu J() = 0. Per saber si s’han de modificar a l’alça o a la baixa els paràmetres o si s’ha arribat a un mínim s’utilitza la derivada. També existeix una variant per la qual hi ha un coeficient que accelera o desaccelera la modificació dels paràmetres.

Un dels problemes en que ens podem trobar es caure en l’error de trobar un mínim local i pensar que es un mínim global, per això es fan diferents execucions de l’algorisme amb inicialitzacions diferents, fins trobar el millor valor de la funció de cost. Els coeficients que tingui en aquest moment seran els millors per a la funció de regressió.

S’explicarà com s’ha programat aquest algorisme en base a les funcions de les que es composa.

1. Inicialització: es passa una llista de valors inicials (els atributs), el paràmetre alpha, el conjunt d’entrenament i el conjunt de test corresponent.
2. Train: és la funció principal i tracta de cridar a la funció predict i després a la funció update. Per paràmetre se li passa max\_iter, que es el màxim d’iteracions que volem que faci el nostre bucle.
3. Predict: aquesta funció prediu la nota G3 a partir dels paràmetres W’s, que anem modificant a cada iteració.
4. Update: s’encarrega d’actualitzar els valors de les W’s calculant l’error entre la predicció de l’atribut objectiu i el valor real d’aquest (el groundtruth), mitjançant la mitjana de l’error quadràtic. Aquest càlcul es el que ens serveix per a calcular el valor de la funció de cost amb aquests paràmetres i actualitzar-los en funció d’aquesta.

Després d’executar l’algorisme amb diferents paràmetres s’ha decidit el següent:

* La inicialització dels pesos es farà de forma aleatòria en un rang de [0.001, 0.500].
* S’executarà la funció train amb un max\_iter de 1000 i una epsilon de 0.00001. Creiem que aquests paràmetres són adequats ja que amb aquests normalment l’algorisme convergeix en menys iteracions amb un error bastant acceptable, tenint en compte només l’epsilon com a criteri de parada. I, en les ocasions en que no convergeix en 1000 iteracions, l’error assolit també és bo, ja que el valor de epsilon es petit.
* L’alpha que dóna millors resultats amb aquest número d’iteracions es 0.0005.

S’ha realitzat la visualització de l’error assolit en funció del número d’iteració en una de les execucions i obtenim la següent gràfica:

