

# Aprendentatge computacional - MatCAD

## Pràctica 1: Regressió

Roger Montané Güell (1569031)

Iker Soto Picón (1565939)

Ignacio Cobas (1571458)

Desembre 2021

### 1 Introducció

En aquesta primera pràctica de l'assignatura sobre Aprendentatge computacional tindrem un primer contacte amb la manera de treballar i analitzar una base de dades per tal d'extreure'n informació útil. Es començarà amb un estudi del data-set donat, explorant-ne els atributs, el significat dels mateixos, els seus valors i fixant quin serà el nostre atribut objectiu. Un cop finalitzat aquest estudi sobre les nostres dades estarem llestos per començar a fer regressió per tal de trobar un model capaç de donar-nos informació valuosa sobre el nostre atribut objectiu a partir de noves dades d'entrada.

Per últim, programarem el mètode del Descens del gradient per tal de comparar-ne els resultats amb els que haurem trobat fent regressió.

Podeu trobar a continuació el [repositori GitHub](#) amb el notebook que conté l'estudi complet.

### 2 Base de dades

La nostra [base de dades](#) conté informació tant per dies com per hores sobre el lloguer de bicicletes de l'empresa [Capital bikeshare](#) a la ciutat de Washington, DC entre els anys 2011 i 2012. Els data-sets contenen les següents columnes (excepte "hr" que només està disponible al data-set per hores):

- **instant:** instant de temps
- **dteday:** data
- **season:** estació de l'any:
  - **1:** primavera

- **2**: estiu
- **3**: tardor
- **4**: hivern
- **yr**: any (0: 2011, 1: 2012)
- **mnth**: mes (1 a 12)
- **hr**: hora (0 a 23)
- **holiday**: 1 si el dia és festiu, 0 sinó
- **weekday**: dia de la setmana
- **workingday**: 1 si el dia no és cap de setmana ni festiu, 0 sinó
- **weathersit**: situació meteorològica:
  - **1**: clar, pocs núvols, parcialment ennuvolat
  - **2**: boira + núvols, boira + pocs núvols, boira
  - **3**: neu lleugera, pluja lleugera, tempesta
  - **4**: pluja intensa + gel + tempesta + boira, neu + boira
- **temp**: temperatura normalitzada en Celsius. Els valors es calculen mitjançant  $\frac{t - t_{min}}{t_{max} - t_{min}}$ ,  $t_{min} = -8$ ,  $t_{max} = 39$  (només en escala horària)
- **atemp**: sensació tèrmica normalitzada en Celsius. Els valors es calculen mitjançant  $\frac{t - t_{min}}{t_{max} - t_{min}}$ ,  $t_{min} = -16$ ,  $t_{max} = 50$  (només en escala horària)
- **hum**: humitat normalitzada. Els valors estan dividits per 100 (el màxim)
- **windspeed**: velocitat del vent normalitzada. Els valors estan dividits per 67 (el màxim)
- **casual**: nombre d'usuaris casuais
- **registered**: nombre d'usuaris registrats
- **cnt**: nombre total de bicis llogades (per usuaris "casual" i "registered")

Amb una primera ullada sobre les dades, és prou trivial veure que un fort candidat a atribut objectiu seria **cnt**, ja que així obtindríem un model capaç de predir el nombre d'usuaris en un cert dia (o en una certa hora) a partir de les altres variables, el qual pot ser de gran interès per a l'empresa llogatera. Per altra banda, també podria ser una bona idea estudiar el nombre d'usuaris registrats segons l'època de l'any, ja que això permetria a l'empresa enfocar millor les seves campanyes de publicitat, ofertes, etc. D'aquesta manera, també podríem veure que fer un model per predir els usuaris casuais segons diferents factors meteorològics també podria ser de cert interès per a l'empresa, ja que d'aquesta manera podrien ser capaços de prevenir la manca de disponibilitat de bicicletes

deurat a un increment de l'anterior tipus d'usuaris.

En aquest estudi, però, ens centrarem únicament en estudiar com es comporta la variable `cnt` respecte les altres variables, i veurem si realment tots els atributs aporten la mateixa informació a l'hora de predir el target.

### 3 Anàlisi de les dades

#### 3.1 Relacions entre variables

Comencem agafant com a atribut objectiu `cnt` i mirant com pot estar relacionat amb les altres variables de la base de dades.

Per començar, sembla intuïtiu que el nombre de bicis llogades hauria d'estar relacionat amb la sensació tèrmica, vegem a la Figura 1 si això es compleix i en cas afirmatiu, vegem com es comporten una variable respecte l'altra.

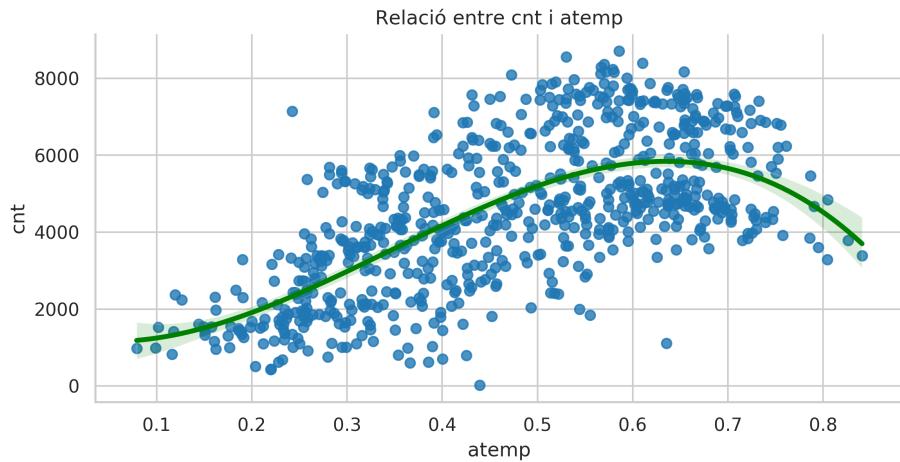


Figure 1: Relació entre `cnt` i `atemp` amb model de regressió i banda de confiança

Observem que quan la sensació tèrmica augmenta, el número de bicis llogades va augmentant fins a cert punt on comença a disminuir. Podem deduir que amb temperatures baixes o molt elevades tindrem menys bicis llogades. Si fem els càlculs de la sensació tèrmica, segons la nostra funció, podem observar que per sota d'aproximadament els 20°C, el lloguer de bicis estaria per sota de les 4000 unitats, i per sobre dels 40 també.

Llavors la temperatura s'hauria de comportar de forma semblant, Figura 2.

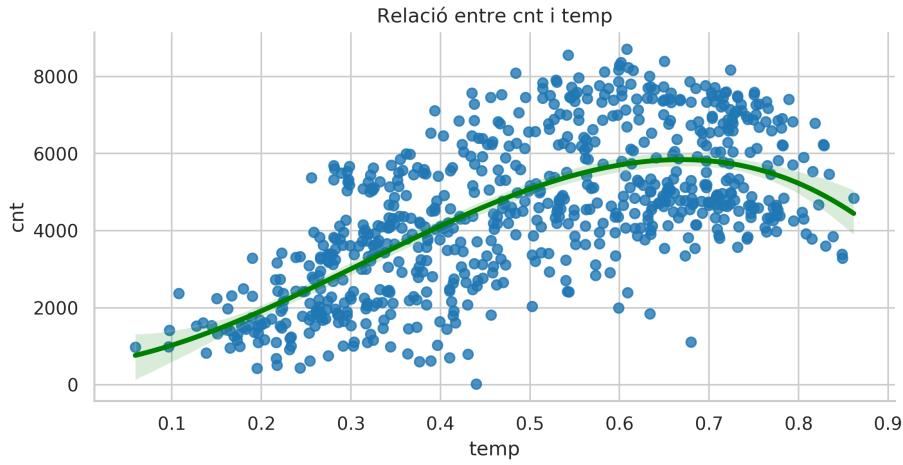


Figure 2: Relació entre cnt i temp amb model de regressió i banda de confiança

És molt semblant a la de la sensació tèrmica ja que la temperatura i aquesta estan relacionades. Si creem un model per aquests podríem esperar trobar-nos una recta amb pendent positiu. També esperem per tant una correlació positiva i alta. Com veiem en la Figura 3, es compleix la nostra hipòtesi i observem també que hi ha un valor molt llunyà a la resta que podem considerar com un outlier.

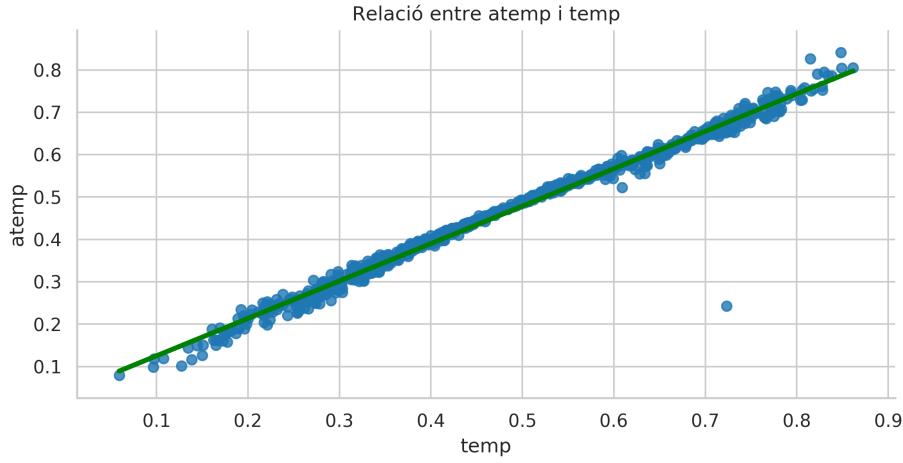


Figure 3: Relació entre atemp i temp amb model de regressió i banda de confiança

Seguint amb l'estudi del lloguer de bicletes respecte les condicions meteorològiques, veiem a continuació com es comporta `cnt` segons la velocitat del vent. Tal i com s'observa a la Figura 4, el nombre de bicletes llogades sembla ser inversament proporcional a la velocitat del vent.

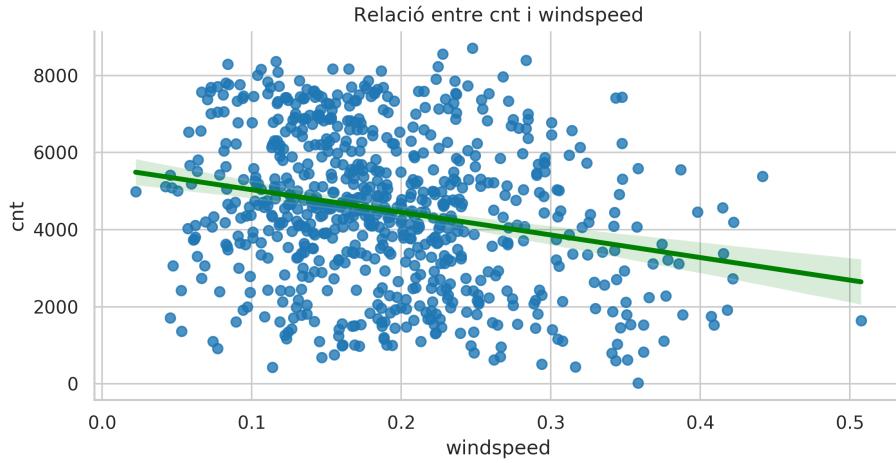


Figure 4: Relació entre cnt i windspeed amb model de regressió i banda de confiança

També observem en la Figura 5 el lloguer de bicis segons l'estació de l'any. Recordem de l'apartat 2 que per aquesta variable tenim:

- 1: primavera
- 2: estiu
- 3: tardor
- 4: hivern

Amb el que observem que l'estiu i la tardor son els mesos on més bicis es lloguen, sent la última la que més, possiblement per ser una època de l'any en que la gent treballa i encara no fa prou fred. En canvi per a la primavera el lloguer de bicis és molt baix, el qual podria ser causat pel fet que hi hagi més dies de pluja en aquell mes.

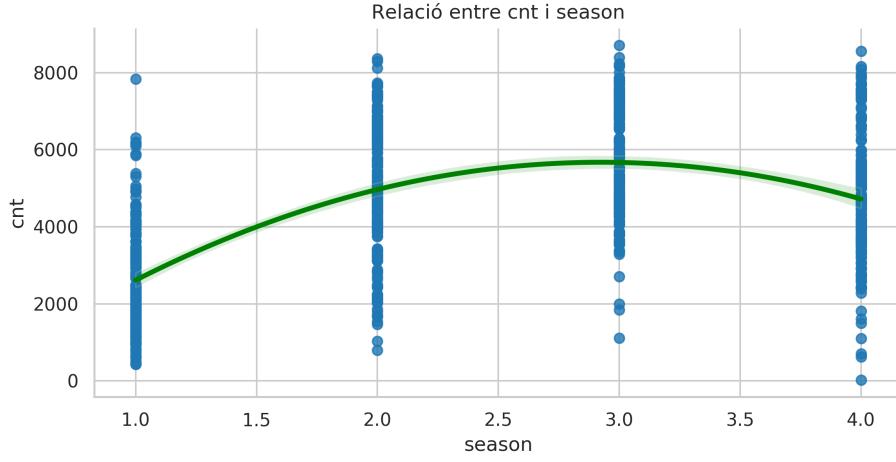


Figure 5: Relació entre cnt i season amb model de regressió i banda de confiança

Per últim, en la Figura 6 observem com es comporta la variables "cnt" respecte l'instant de temps (i.e. vegem el nombre de bicis llogades cada dia).

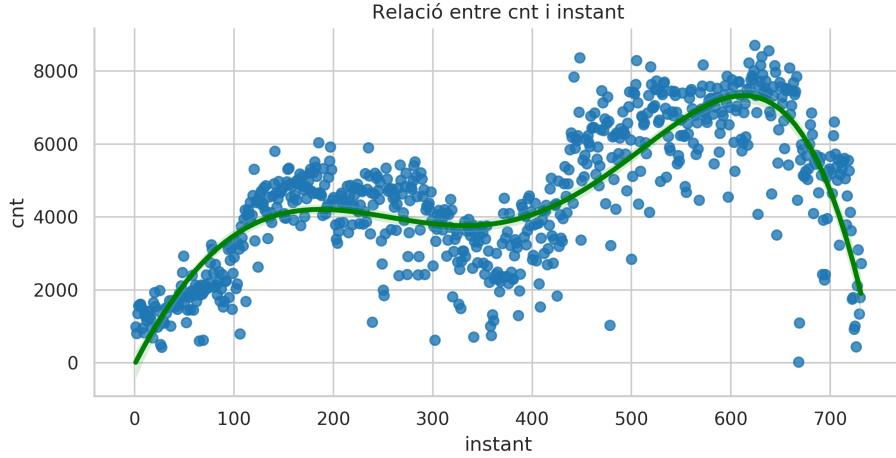


Figure 6: Relació entre cnt i instant amb model de regressió i banda de confiança

S'observa que hi ha una mena de cicle que es repeteix cada any: al gener el lloguer de bicis és més baix, s'assoleix un màxim de lloguers a mitjans d'any i el nombre de lloguers torna a baixar a mesura que torna el fred. Això sembla repetir-se el 2012 però donat que l'empresa haurà guanyat popularitat durant el primer any, al gener del segon any comença amb un mínim d'aproximadament

4000 lloguers i el màxim (assolit aproximadament passat mig any) supera els 8000 lloguers en un dia. Veiem però que un cop torna a arribar el fred el nombre de lloguers decau fins per sota de 2000 al dia.

### 3.2 Histogrames

Com s'ha comentat en l'apartat 2, la base de dades consta de 2 fitxers, un en que tenim el lloguer de bicicletes per dia (i l'utilitzat en aquest estudi) i un en que tenim el lloguer de bicicletes per hora. D'aquest últim podem treure un histograma (Figura 7) que ens mostri el lloguer de bicis total de bicis a cada hora del dia, on s'observa que tant a les 8 del matí com a les 5 de la tarda hi tenim els pics més alts, que sembla coincidir amb l'hora d'entrada i sortida de la feina o escola de la majoria de la població.

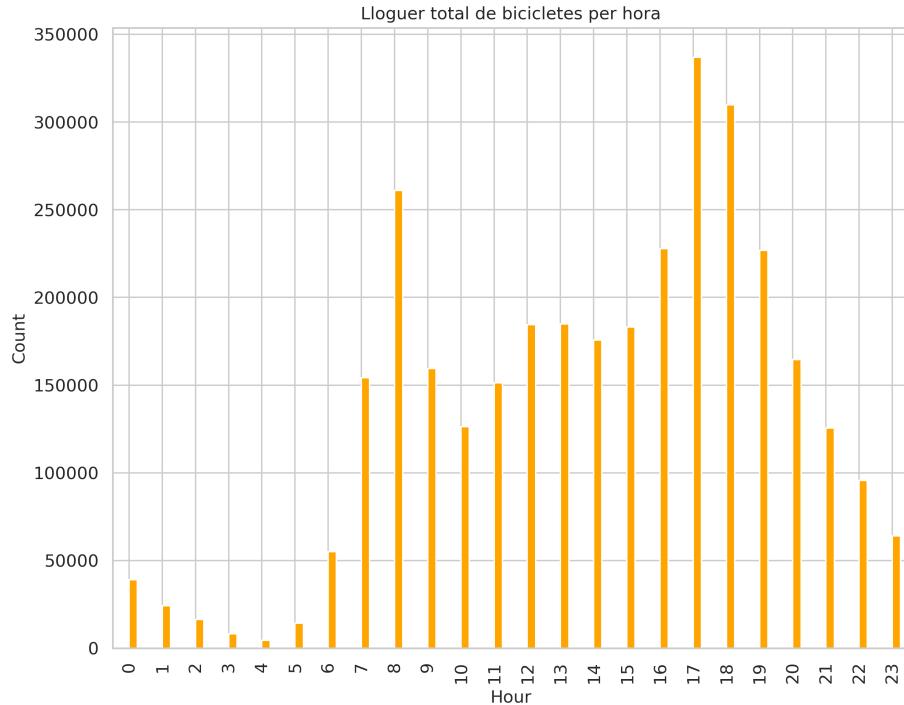


Figure 7: Histograma del lloguer total de bicicletes per hora

A més, també podem veure el lloguer total per hora en dies laborables en la Figura 8, el qual manté una forma semblant a l'anterior (els pics són els mateixos) però amb valors més baixos.

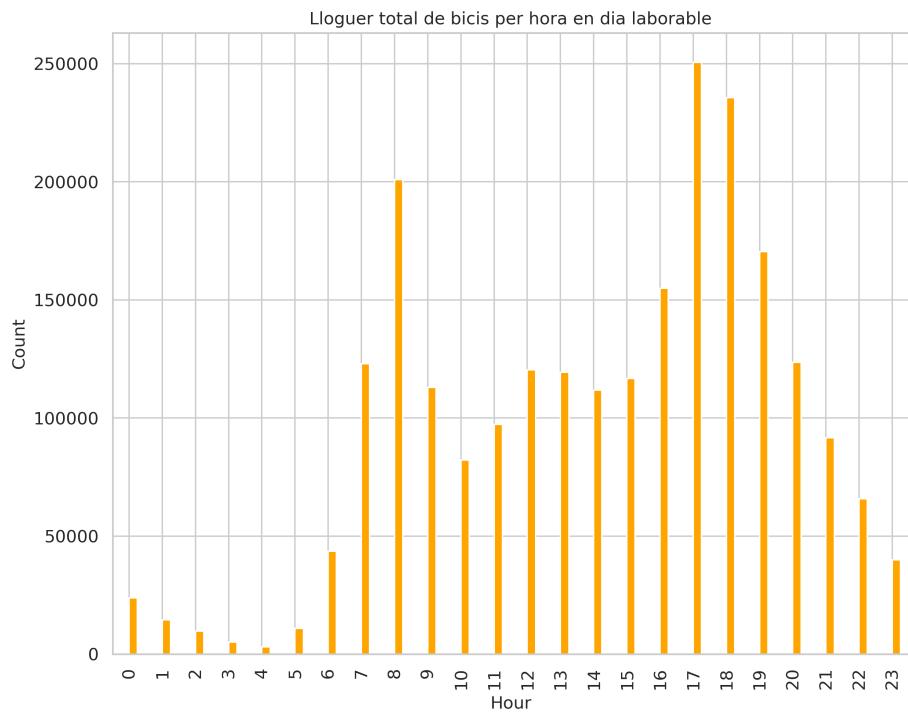


Figure 8: Histograma del lloguer total de bicicletes per hora en dies laborables

En quant als caps de setmana (Figura 9) però, tot i que els pics es mantenen a les mateixes hores, no són tant pronunciats i el nombre de bicis llogades entre les 11 del matí i les 4 de la tarda és molt més alt, inclús superior al pic de les 8 del matí.

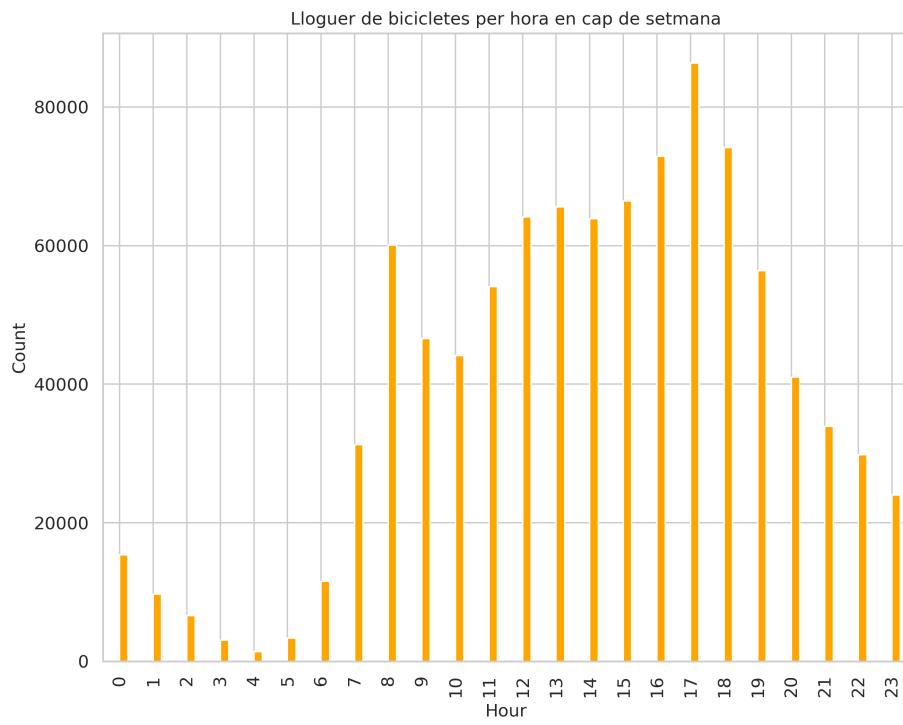


Figure 9: Histograma del lloguer total de bicicletes per hora en caps de setmana

Com hem vist en l'apartat 3.1, el major nombre de bicis es lloguen durant l'estiu i la tardor, ara bé, observem en la Figura 10 també que la gran majoria de bicis es lloguen quan les condicions meteorològiques són bones ('clear'). Quan hi ha boira lleugera o molta neu també se'n lloguen però molt poques i quan hi ha molta pluja no se'n lloguen.

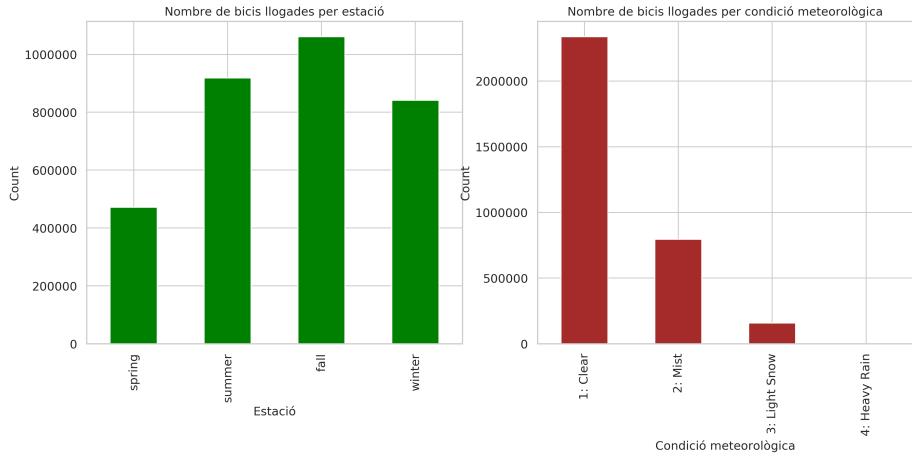


Figure 10: Histogrames del lloguer total de bicis segons l'estació de l'any (a l'esquerra) i segons les condicions meteorològiques (a la dreta)

Mirem ara l'histograma de l'atribut `cnt` (Figura 11).

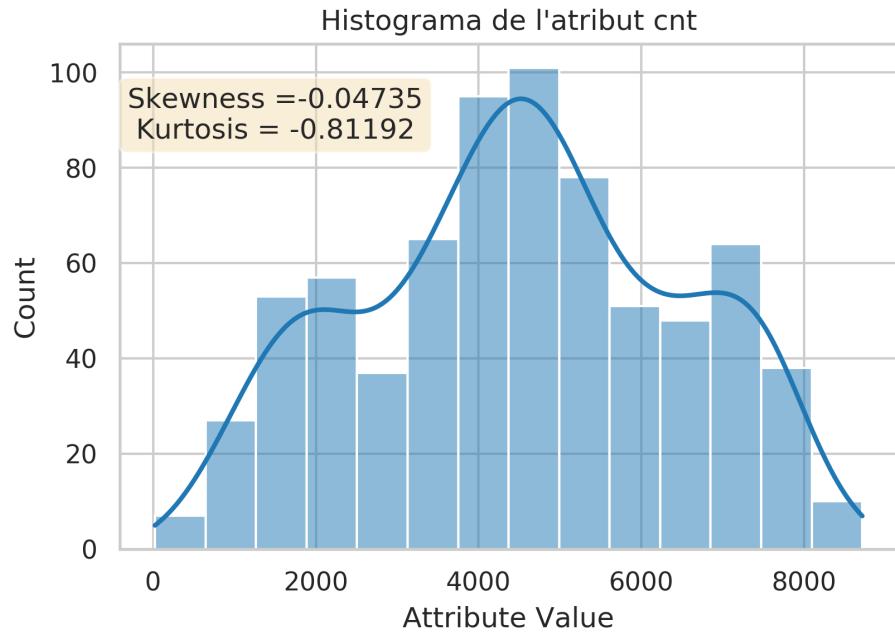


Figure 11: Histograma de l'atribut `cnt`

### 3.3 Correlació entre variables

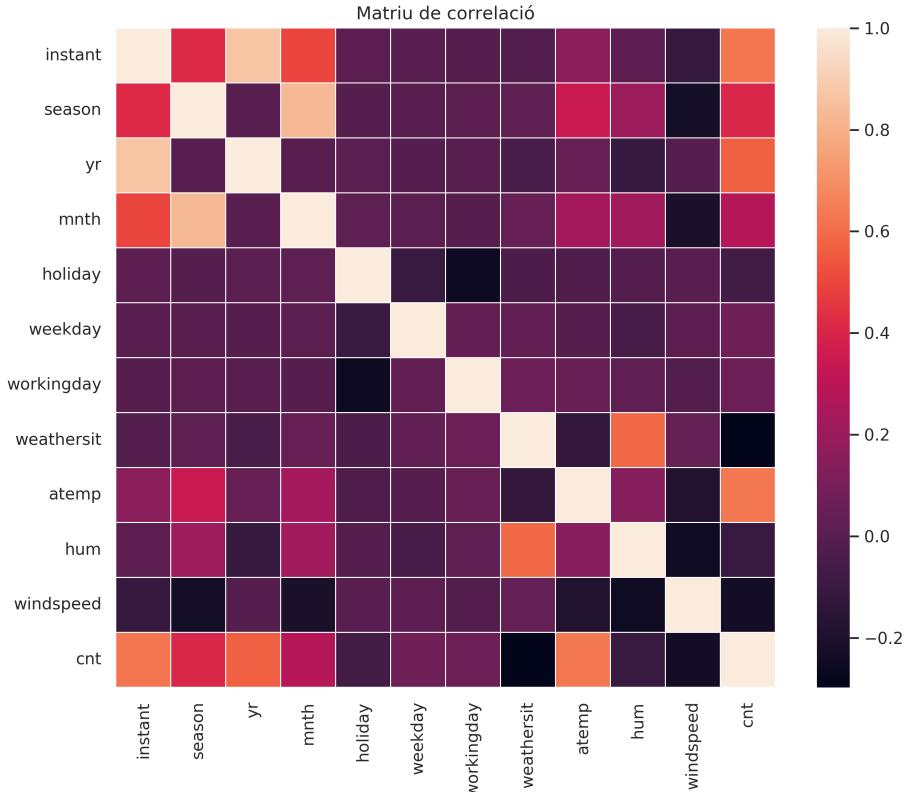


Figure 12: Matriu de correlació

Mirant la matriu de correlació en la Figura 12 veiem que la majoria de variables no tenen correlació entre si. Si ens fixem en la variable "cnt", veiem que està molt correlacionada amb la variable "registered" el qual ens podrem esperar, però a més té una correlació aproximadament major a 0.6 amb les variables "casual", "atemp", "temp", "yr" i "instant". Més endavant veurem si aquestes variables seràn les millors per als models de regressió.

## 4 Regressió

Per tal de ser capaços de predir el nostre target `cnt` farem regressió lineal i polinomial.

Abans de començar amb la regressió, però, ens caldrà primer estandarditzar

les dades. Si un conjunt de dades  $X$  arbitrari amb mitja  $\bar{x}$  i variació estàndard  $\sigma$ , per tal d'obtenir el conjunt estandarditzat  $\hat{X}$ :

$$\hat{x}_i = \frac{x_i - \bar{x}}{\sigma}$$

Aquestes noves dades estandarditzades les podem utilitzar per veure quines de les variables estan distribuïdes segons la Normal (o Gaussiana) ja que seran les variables preferides per a fer la regressió. A més, també haurem de descartar certes variables que no són representatives per a la regressió i que l'únic que fan és afegir soroll al model.

Un cop generat el nou conjunt  $\hat{X}$  ja podem començar a fer regressió. Fent servir la llibreria Sklearn fem una primera regressió lineal de la qual en veiem el resultat a la Figura 13. Aquesta regressió ens genera un Error Quadràtic Mitjà (o MSE en anglès) d'aproximadament 0.6 i un  $R^2$  d'aproximadament 0.4. Aquests resultats, però, només tenen en compte la relació entre el target `cnt` i la variable `instant`, i el que ens interessa és fer regressió multi-dimensional de manera que trobem les relacions entre les variables del dataset i el target.

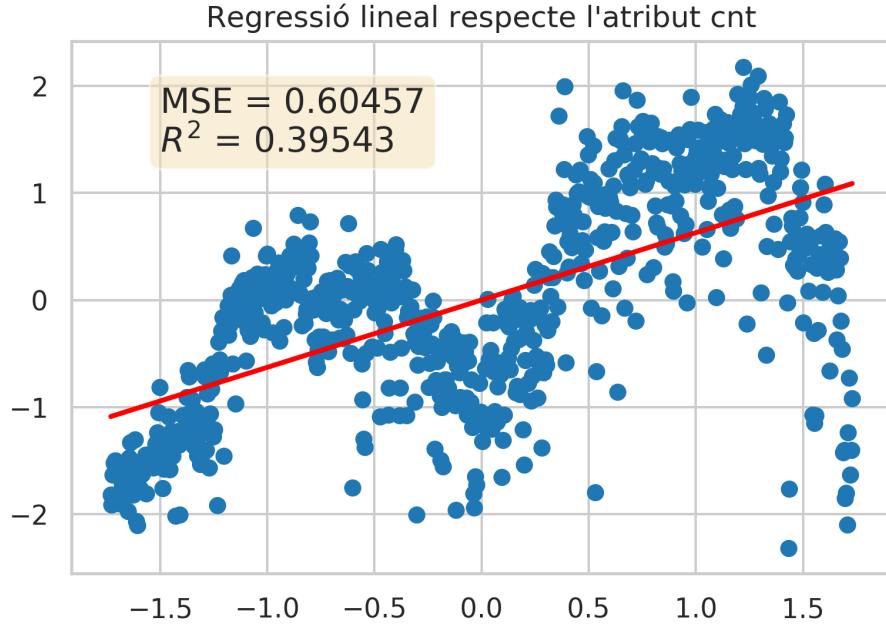


Figure 13: Resultat de la regressió lineal amb un sol atribut

Ara, veiem que podem avaluar de manera independent com d'idoni és cada

atribut. Això ens ajudarà a fer la tria d'aquells atributs més rellevants per a la regressió, descartant aquells que no siguin rellevants o que simplement acaben generant soroll per al model.

```
Error en instant: 0.380860
R2 score en instant: 0.550589
Error en season: 0.685485
R2 score en season: 0.191134
Error en yr: 0.481432
R2 score en yr: 0.431914
Error en mnth: 0.761463
R2 score en mnth: 0.101480
Error en holiday: 0.854338
R2 score en holiday: -0.008111
Error en weekday: 0.898483
R2 score en weekday: -0.060202
Error en workingday: 0.853487
R2 score en workingday: -0.007107
Error en weathersit: 0.828772
R2 score en weathersit: 0.022056
Error en atemp: 0.511307
R2 score en atemp: 0.396663
Error en hum: 0.881572
R2 score en hum: -0.040247
Error en windspeed: 0.801444
R2 score en windspeed: 0.054303
```

Figure 14: Avaluació individual dels atributs

A més, volem veure també com s'ajusta el model a noves dades d'entrada, per tant, separarem les nostres dades de manera que un 80% s'utilitzin per a l'entrenament i un 20% per a la validació.

## 4.1 Regressió lineal

Per als tres regressors següents, començarem per generar un model amb tots els atributs per després comprar-ne els resultats generant un nou model amb només els  $k$  atributs més importants.

A més, un altre mètode per simplificar el model que se sol utilitzar és el PCA. El provarem i comprovarem si el millor és utilitzar tots els components, només els que hem considerat més importants, o potser transformar-los amb un PCA.

Per al PCA també ens caldrà mirar quin és el millor nombre de components en el qual volem transformar la  $X$ .

### 4.1.1 Regressor lineal

Comencem fent regressió lineal amb tots els atributs. Veiem la Figura 15.

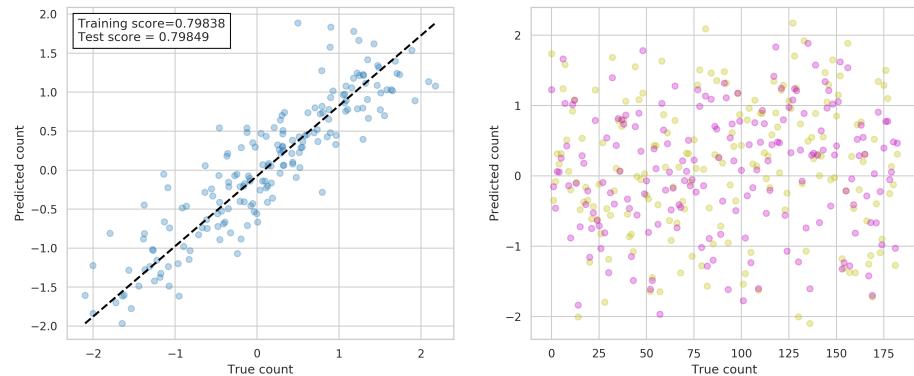


Figure 15: Regressió lineal amb tots els atributs

Veiem que els resultats s'ajusten bastant a la recta diagonal del gràfic a l'esquerra (si el regressor fos perfecte, els punts s'ajustarien tots sobre la recta). A més, a la dreta veiem que tot i que la majoria de les prediccions (en lila) no són perfectes, s'aproximen bastant als valors reals (en groc). Veiem per tant que tant el training com el test score es troben al voltant del 80% i per tant tenim un regressor bastant bo.

Ara, ordenem els atributs de més importants a menys i en seleccionem els  $k$  més importants. Primer haurem de trobar aquesta  $k$ . Com veiem a la Figura 16 obtenim que el millor valor per a  $k$  serà  $k = 6$ .

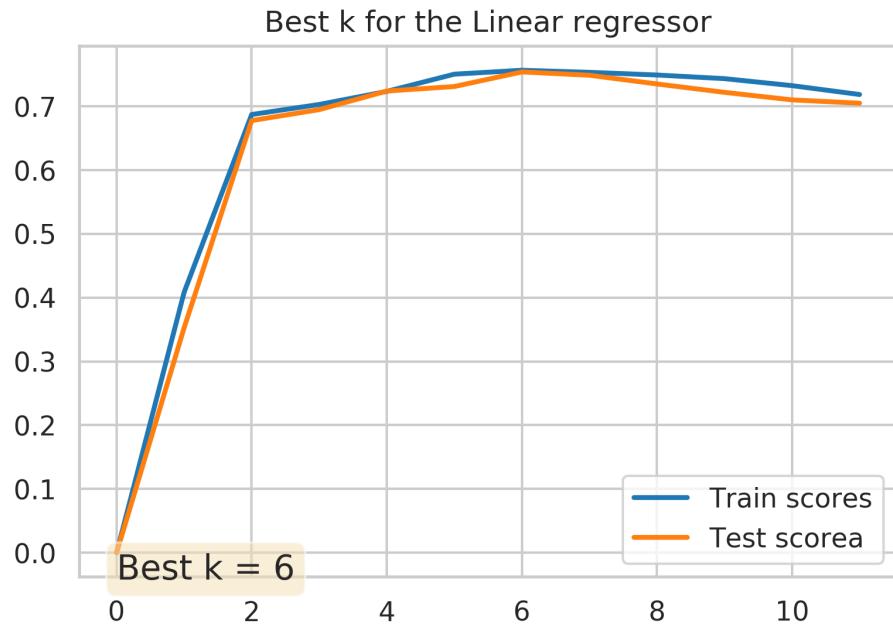


Figure 16: Millor k pel regressor lineal

Cal notar que per a trobar aquesta constant hem "castigat" la complexitat del model de manera que al seu score se li resta:

$$\lambda \left( \frac{k}{k_{max}} \right)^2$$

Agafant els  $k$  millors atributs doncs, obtenim (Figura 17):

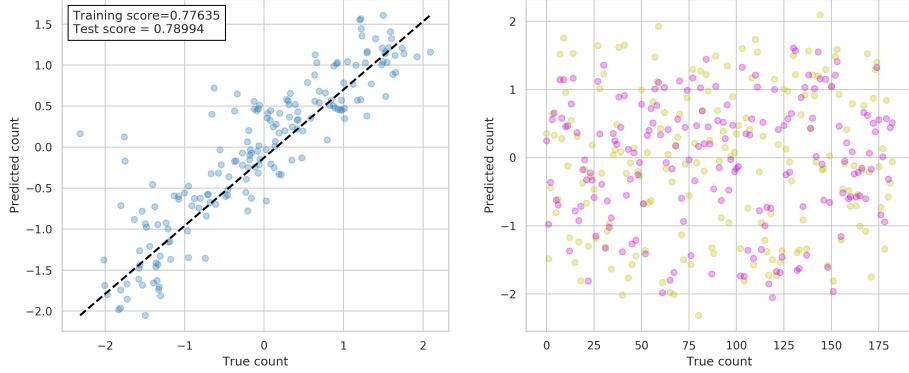


Figure 17: Resultat de la regressió lineal amb les millors variables

Els resultats són bastant semblants al regressor amb totes les variables (inclusivament pitjors) ja que tenim uns scores al voltant del 80%.

Per últim, passem a simplificar el model mitjançant PCA. De nou, haurem de trobar el millor nombre de components en què transformar el conjunt de dades  $X$ , i per al regressor lineal tenim que el millor nombre de components és 8. Un cop transformat, obtenim (Figura 18):

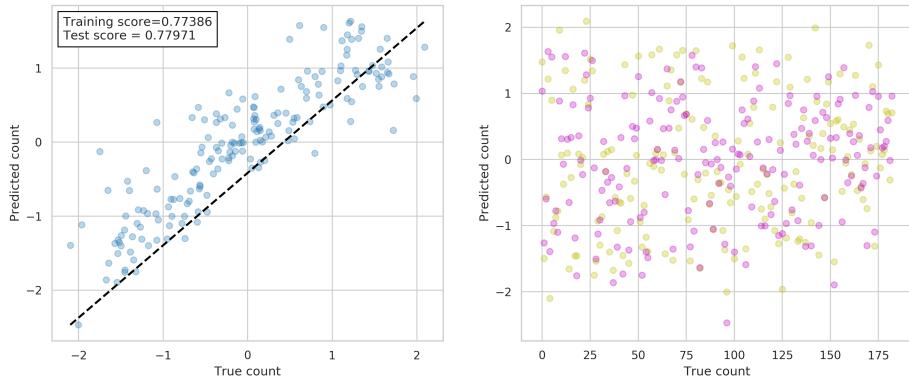


Figure 18: Resultat de la regressió lineal amb PCA

Els resultats són de nou molt semblants als dos anteriors, obtenint un regressor amb un score al voltant del 80%, tot i que en els dos últims casos els punts semblen estar una mica més dispersos de la recta.

#### 4.1.2 Regressor Lasso

Passem ara a generar un Sparse model amb un regressor Lasso per tal d'intentar millorar el nostre model.

També observem que si utilitzéssim les dades sense estandarditzar per al regressor Lasso els scores són més baixos tant per training com testing, i els MSE són molt elevats, per això és important normalitzar les dades.

Observem els resultats obtinguts a la Figura 19.

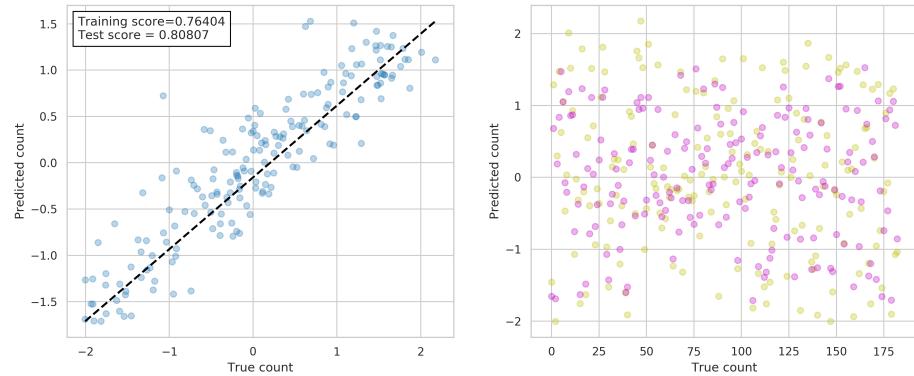


Figure 19: Resultat de la regressió Lasso amb totes les variables

Aquests resultats són bastant semblants als que hem obtingut amb el regressor anterior, però és pitjor, els score es troben al voltant del 75%.

Com abans, ordenem els atributs de més importants a menys i en seleccionem els  $k$  més importants. Veiem a la millor  $k$  a la Figura 20 obtenim que el millor valor per a  $k$  serà  $k = 6$ .

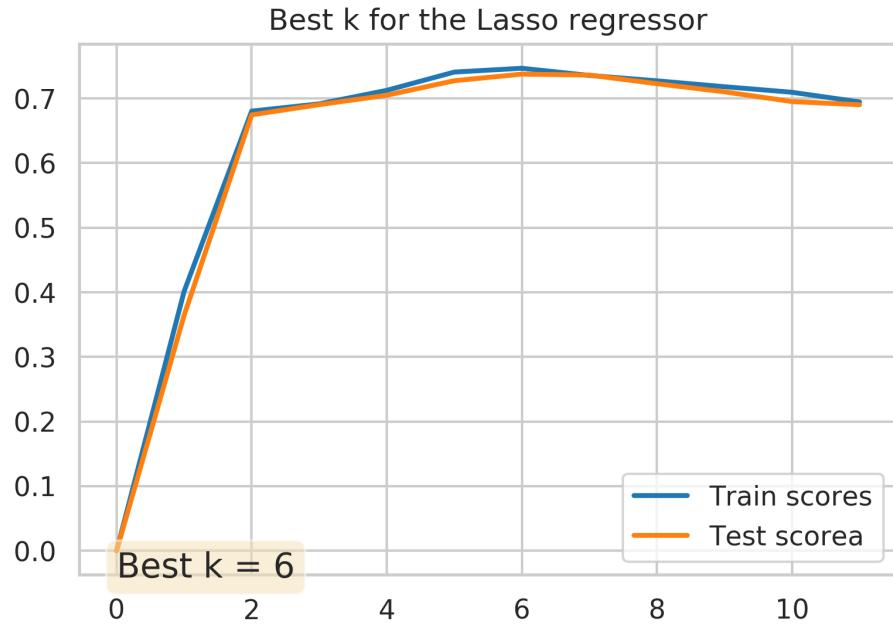


Figure 20: Millor k pel regressor Lasso

Agafant els  $k$  millors atributs doncs, obtenim (Figura 21):

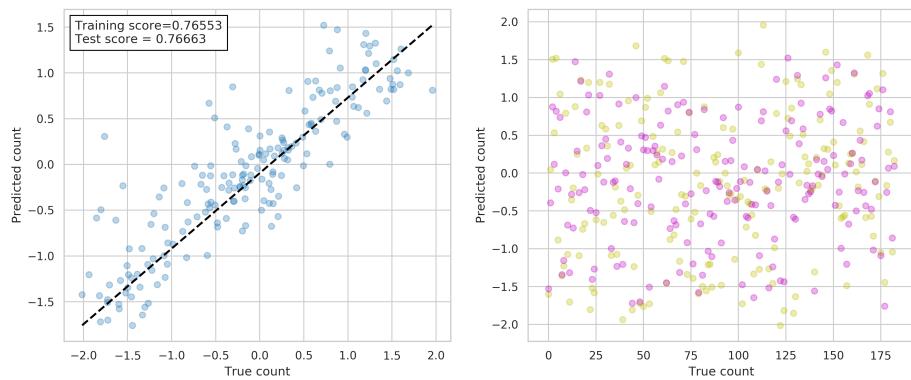


Figure 21: Resultat de la regressió Lasso amb les millors variables

Els resultats són bastant semblants al regressor amb totes les variables, una mica millor, ja que tenim uns scores al voltant del 80%.

Per últim, passem a a simplificar al model mitjançant PCA. De nou, haurem de trobar el millor nombre de components en que transformar el conjunt de dades  $X$ , i per al regressor lineal tenim que el millor nombre de components és 9. Un cop transformat, obtenim (Figura 22):

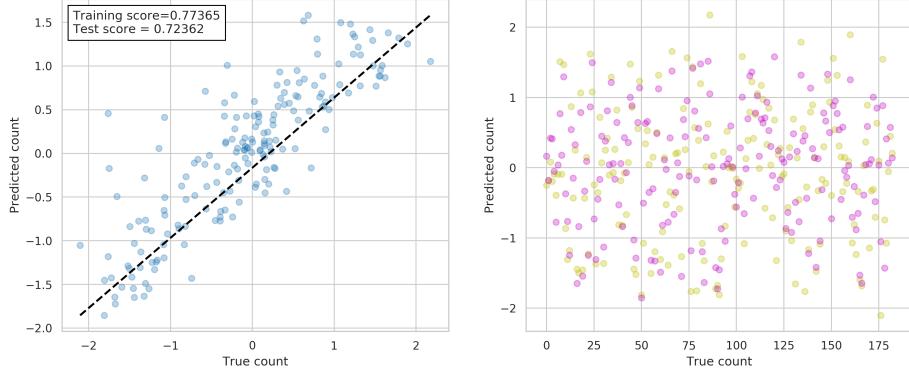


Figure 22: Resultat de la regressió Lasso amb PCA

Els resultats han ara empitjorat, obtenint un regressor amb un score d'entrenament sobre el 75% i el del test al voltant del 70%.

#### 4.1.3 Regressor Ridge

Anem ara a utilitzar la regressió Ridge en la qual els coeficients calculats minimitzen la suma dels quadrats dels residus penalitzada al afegir el quadrat de la norma L2 del vector format pels coeficients. Mostrem a la Figura 23 doncs els resultats de fer la regressió Ridge amb totes les variables:

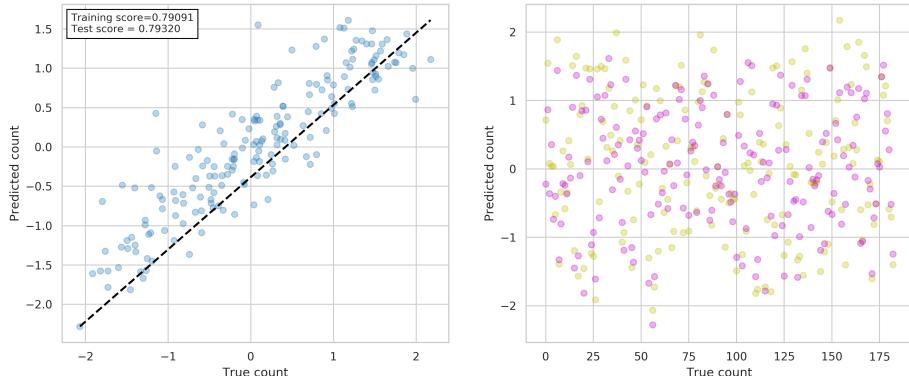


Figure 23: Resultat de la regressió Ridge amb totes variables

Amb les millors  $k$  variables (mostrant primer el millor valor de  $k$ ) en les Figures 24 i 25 respectivament:



Figure 24: Millor valor de  $k$  per a la regressió Ridge

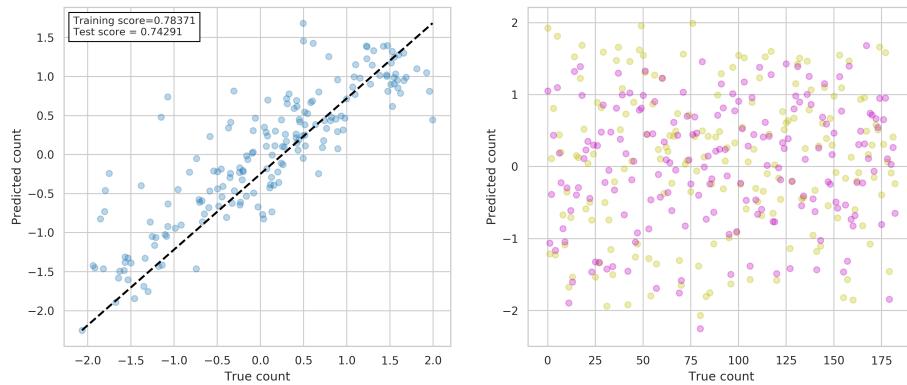


Figure 25: Resultat de la regressió Ridge amb les millors variables

I amb PCA en la Figura 26:

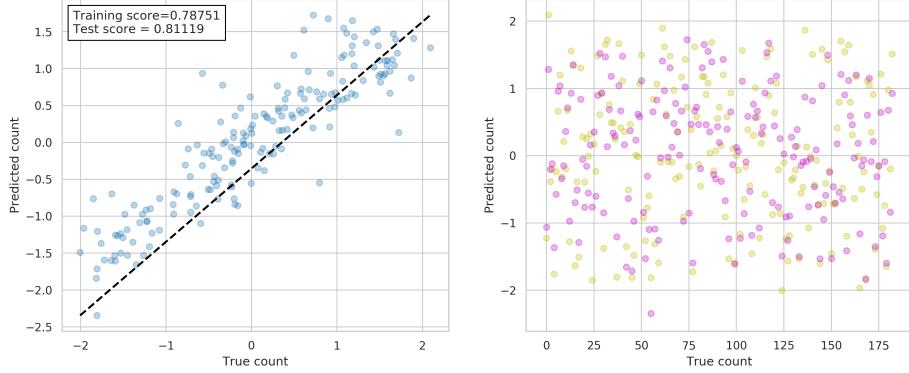


Figure 26: Resultat de la regressió Ridge amb PCA

Veiem que, de nou, els tres regressors tenen uns scores al voltant del 80%, sent el millor els regressor amb PCA que obté un score de test del 81.19%. En els tres casos a més, els punts es distribueixen de forma bastant semblant sobre la recta, tot i que en el cas en que usem les  $k$  millors variables semblen estar més dispersats.

## 4.2 Regressió polinomial

Tot i estar obtenint uns scores que podriem considerar com a prou bons, podria ser que els regressor lineals no puguin donar més de si per aquest problema i ens calgui anar més enllà. Provem doncs de fer un model de regressió polinomial. Primer de tot haurem de trobar quin és el grau més adient del polinomi aproximador. Ho fem generant diversos models amb tots els graus possibles i acabem obtenint que el millor és grau 2.

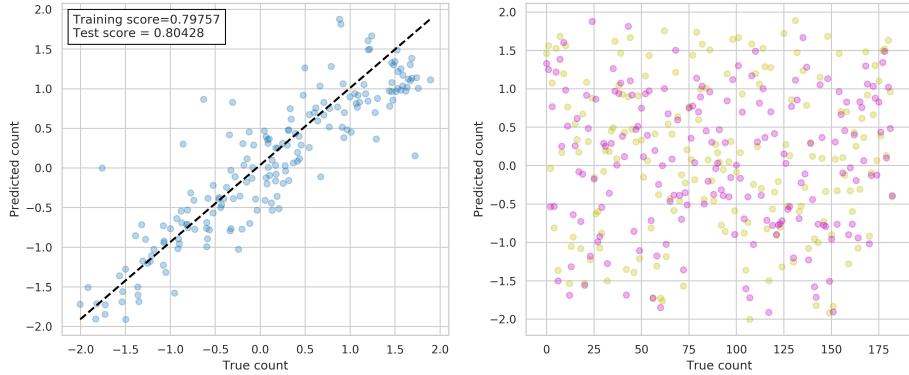


Figure 27: Resultat de la regressió polinomial amb tots els atributs

Com veiem a la Figura 27, obtenim uns scores més bons que per a la regressió lineal i sense fer cap transformació de les dades (a part de la normalització). Amb la crossvalidation que hem fet per trobar el millor grau obtenim un test score per sobre del 85% el qual no hem vist en cap dels regressors lineals.

### 4.3 Altres regressors

Per últim, provarem ràpidament tres models que creiem que amb poques dades (menys de 1000) poden funcionar bé.

#### 4.3.1 Random forest

Comencem per un Random forest, observem els resultats en la Figura 28.

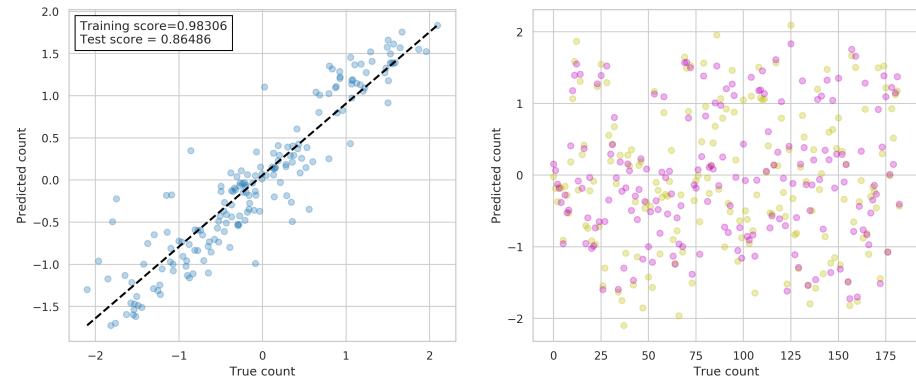


Figure 28: Resultat de la regressió amb Random forest i tots els atributs

Per aquest model el training score augmenta molt situant-se al 98% mentre que el test score es manté bastant semblant al cas del regressor polinomial, per sobre del 85%.

#### 4.3.2 Regressor d'arbre de decisió

A continuació provarem un Decision tree regressor, obtenim els resultats a la Figura 29.

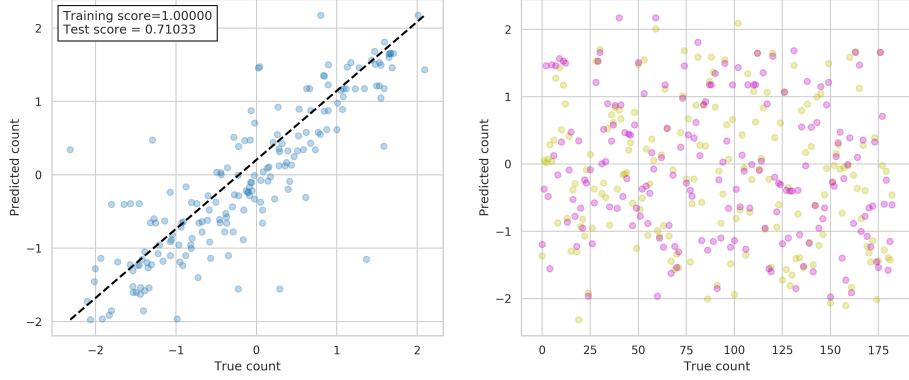


Figure 29: Resultat de la regressió amb Decision tree classifier i tots els atributs

Veiem que si be el score de l'entrenament és del 100% el del test és del 71%, més baix que els scores obtinguts en els regressors lineals.

#### 4.3.3 Gradient boosting

Acabarem amb un model de Gradient boosting en la Figura 30:

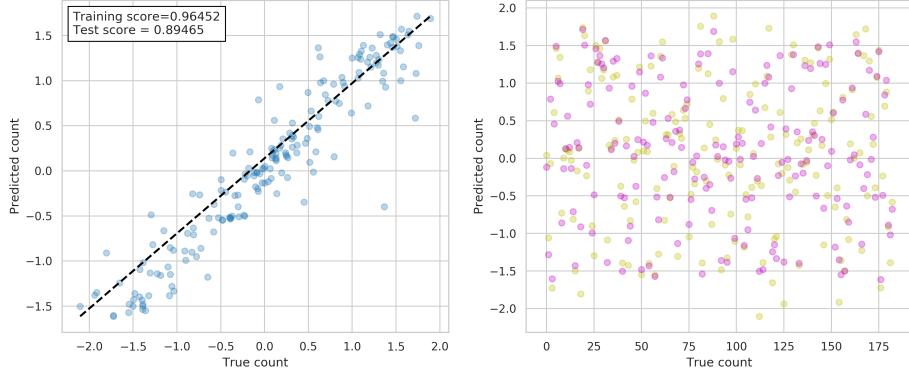


Figure 30: Resultat de la regressió amb Gradient boosting i tots els atributs

Com veiem en la Figura 30 el score de l'entrenament és del 96% i el del test és del 89% que és el més alt de tots els models. Observem que els punts estan distribuïts de manera bastant uniforme sobre la recta, a excepció d'alguns.

## 5 El descens del gradient

En aquest apartat, ens centrarem en implementar en python el procés de descent del gradient explicat a les classes de teoria, i comparar-lo amb els resultats obtinguts amb el regressor.

$$J(w) = \frac{1}{2m} \left[ \sum_{i=1}^m (f(x^i; w) - y^i)^2 + \lambda \sum_{j=1}^n (w_j^2) \right]$$

*J* retorna el `mse`. Per a trobar  $w_j$ , repetirem fins convergència:

$$w_0 = w_0 - \alpha \frac{1}{m} \sum_{i=1}^m (f(x^i; w) - y^i) \cdot 1$$

$$w_j = w_j - \alpha \left[ \frac{1}{m} \sum_{i=1}^m (f(x^i; w) - y^i) \cdot x_j^i - \frac{\lambda}{m} w_j \right]$$

ó:

$$w_j := w_j \left( 1 - \alpha \frac{\lambda}{m} \right) - \alpha \frac{\lambda}{m} \sum_{i=1}^m (f(x^i; w) - y^i) \cdot x_j^i$$

On si considerem un regressor lineal (el model és una recta), llavors  $w_0$  i  $w_1$  representen, respectivament, la  $b$  i  $a$  de la fórmula de la recta:

$$h_\theta(x^{(i)}) = ax + b$$

$\alpha$  és el learning rate, i  $h_\theta(x^{(i)})$  és la funció que fa la regressió, és a dir, la funció que prediu el valor de  $y^{(i)}$  donat un(s) atribut(s) concret(s)  $x^{(i)}$ .

La funció per realitzar el procés és `gradientDescent`, els seus argument són:

- `x`: dades del dataset.
- `y`: objectiu del dataset.
- `theta`: vector que conté els pesos que retornarem.
- `alpha`: learning rate.
- `m`: nombre de mostres.
- `numIterations`: nombre de vegades que iterarem durant el procés.

I retorna el vector `theta` amb els pesos finals.

Una vegada que tenim els pesos de cada atribut podem comparar amb els obtinguts amb el regressor lineal construït prèviament.

Ordenant els atributs de menys a més important segons el regressor lineal

obtenim el següent ordre:

```
[‘instant’, ‘weekday’, ‘workingday’, ‘mnth’, ‘holiday’, ‘season’, ‘weathersit’,  
‘hum’, ‘windspeed’, ‘yr’, ‘atemp’]
```

I amb el descens del gradient:

```
[‘instant’, ‘yr’, ‘holiday’, ‘windspeed’, ‘hum’, ‘atemp’, ‘workingday’,  
‘weathersit’, ‘season’, ‘mnth’, ‘weekday’]
```

Les diferències més notables les veiem en els atributs de temps excepte l'instant, és a dir, en 'yr', 'mnth', 'weekday'. En el primer regressor el dia de la setmana ('weekday') i el mes ('mnth') tenien poc pes i l'any ('yr') era dels més importants, mentre que en el descens del gradient és completament l'oposat. L'any no te gairé pes i els altres dos atributs en tenen força.

## 5.1 Learning rate

El valor del learning rate o la alpha que passem a la funció, representa la mesura dels passos que donem per arribar a convergir. El problema de fer el valor molt petit, és que haurem de fer molts passos i per tant trigarem molt. També tenim problemes si el fem molt gran, ja que el pas pot ser més gran que la diferència d'on som a la solució i saltar-nos aquesta i per tant també augmentaria el temps que amb un learning rate adequat.

El nostre valor pel learning rate hem decidit que sigui 0.005.

## 5.2 Figura 3D del descens del gradient

Aquí creem una figura en 3 dimensions per representar el resultat obtingut al realitzar el descens del gradient, observem el resultat en la Figura 31.

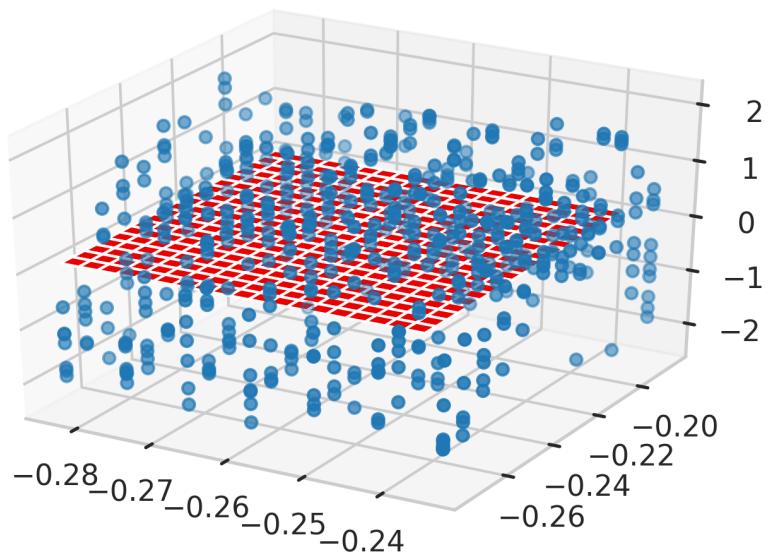


Figure 31: Figura 3D del descens del gradient