Pràctica 1 - Regressió

Aprenentatge Computacional

Enginyeria Informàtica 2021/2022 - UAB

Martí Caixal Joaniquet - 1563587

Ricard López Olivares - 1571136

Jairo Villarroel Rodriguez - 1571069

Introducció

En aquest informe ensenyarem com un anàlisi de regressió ens permet entendre i preveure el comportament de les dades. A través d'unes mostres es fa un aprenentatge dels valors i la seva correlació.

Els objectius generals són els següents:

1. Aplicar models de regressió, posant èmfasi en:
    a. Analitzar els atributs per seleccionar els més representatius i normalitzar-los.
    b. Avaluar correctament l'error del model
    c. Visualitzar les dades i el model resultant
    d. Saber aplicar el procés de descens del gradient
2. Ésser capaç d'aplicar tècniques de regressió en casos reals
3. Validar els resultats en dades reals
4. Fomentar la capacitat per presentar resultats tècnics d'aprenentatge computacional de forma adequada davant altres persones

En aquest cas s'utilitza el Dataset de Boston House Prices. És un seguit de registres que cada un representa una zona de Boston al 1970.

Cada registre està compost per els següents atributs:

- CRIM: tassa de criminalitat per càpita a la zona.
- ZN: proporció de terreny residencial dividit en zones de 25.000 peus quadrats.
- INDUS: proporció de comerços no minoristes per zona.
- CHAS: és una variable per saber si limita amb un riu, (1 limita amb un riu; 0 en cas contrari).
- NOX: concentració d'òxid nítric (part per 10 milions)
- RM: mitjana d'habitacions per habitatge.
- AGE: proporció d'ocupabilitat pels propietaris a partir de 1940.
- DIS: distància ponderada entre 5 llocs de treball
- RAD: índex d'accessibilitat per carretera
- TAX: taxa d'impost per la propietat de cada 10.000 dollars
- PTRATIO: proporció d'alumne-professor per zona
- B: percentatge de persones negres a la zona
- LSTAT: % de persones de baix estatus

- MEDV: mitjana del preu dels habitatges ocupats en 1000s dòlars.

L'objectiu final és poder preveure el valor que tindrà l'atribut "MEDV", és a dir el valor de venda d'una casa. Es decideix aquest i no un altre perquè creiem que és el de major importància i del que es podrien demanar més prediccions

Analitzant les dades

Davant del conjunt de dades que tenim disponibles, primer cal analitzar-les i conèixer quins atributs són els més importants, quins tenen correlació entre ells i quins no són importants.

Primer mostrem quins camps hi ha a la base de dades juntament amb el tipus de dades que contenen. També es mostra que d'un total de 506 registres en total, tots els atributs tenen 506 valors. El fet de no tenir valors nuls simplifica el treball i evita haver de modificar o esborrar entrades no completes.

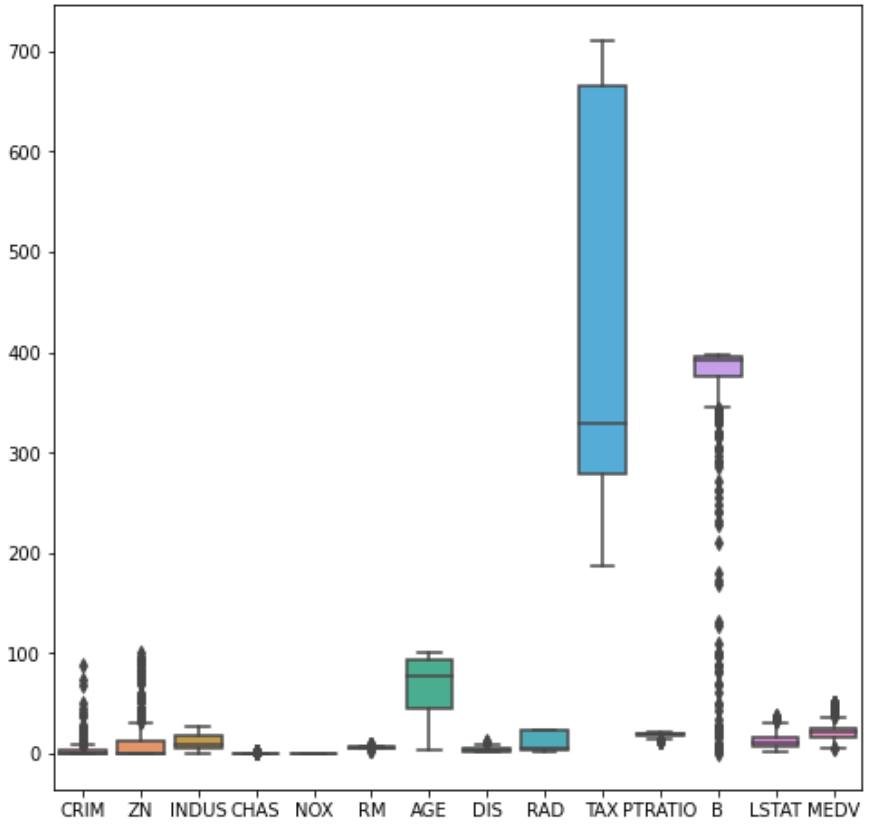
```
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   CRIM     506 non-null    float64
 1   ZN       506 non-null    float64
 2   INDUS    506 non-null    float64
 3   CHAS     506 non-null    int64
 4   NOX      506 non-null    float64
 5   RM       506 non-null    float64
 6   AGE      506 non-null    float64
 7   DIS      506 non-null    float64
 8   RAD      506 non-null    int64
 9   TAX      506 non-null    float64
 10  PTRATIO  506 non-null    float64
 11  B        506 non-null    float64
 12  LSTAT    506 non-null    float64
 13  MEDV     506 non-null    float64
```

Per fer-nos una idea de les dades amb les que treballem, mostrem les 5 primeres entrades del dataset.

Anotació: Les columnes de la taula són cada un dels registres, mentre que les files són els atributs de cada registre

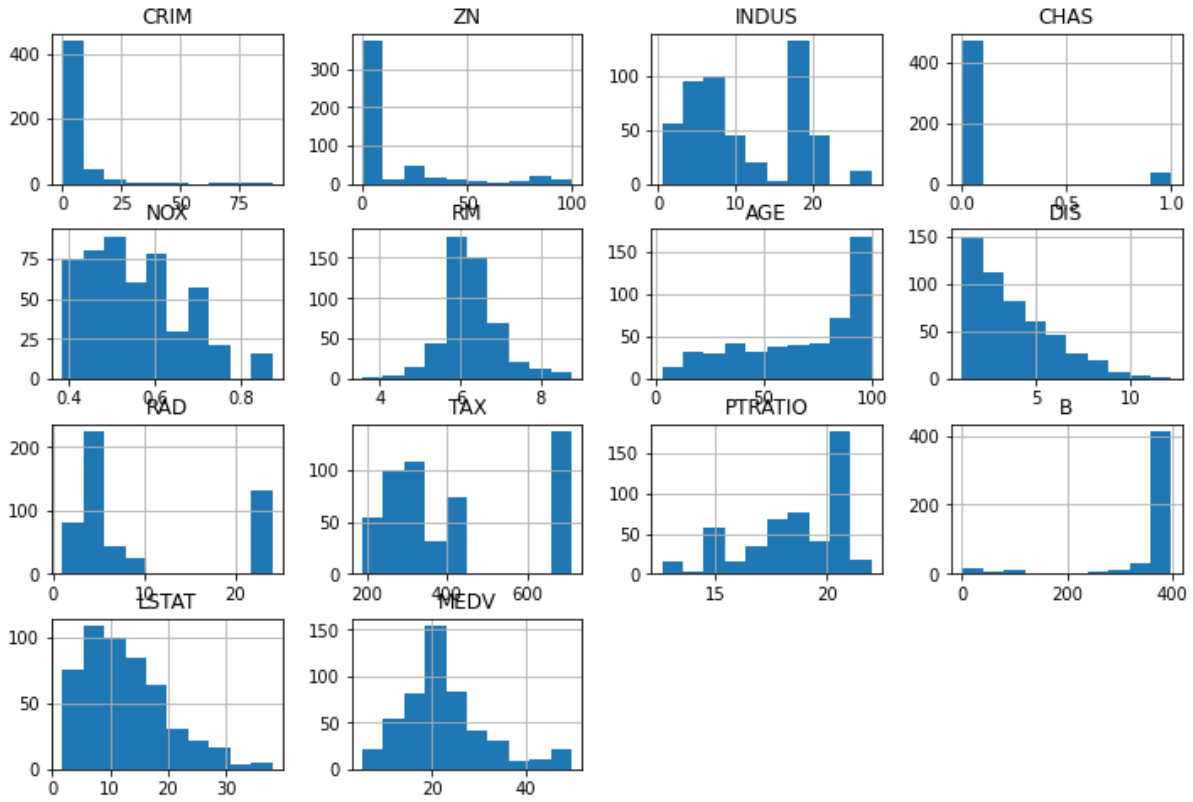
	0	1	2	3	4
CRIM	0.00632	0.02731	0.02729	0.03237	0.06905
ZN	18.00000	0.00000	0.00000	0.00000	0.00000
INDUS	2.31000	7.07000	7.07000	2.18000	2.18000
CHAS	0.00000	0.00000	0.00000	0.00000	0.00000
NOX	0.53800	0.46900	0.46900	0.45800	0.45800
RM	6.57500	6.42100	7.18500	6.99800	7.14700
AGE	65.20000	78.90000	61.10000	45.80000	54.20000
DIS	4.09000	4.96710	4.96710	6.06220	6.06220
RAD	1.00000	2.00000	2.00000	3.00000	3.00000
TAX	296.00000	242.00000	242.00000	222.00000	222.00000
PTRATIO	15.30000	17.80000	17.80000	18.70000	18.70000
B	396.90000	396.90000	392.83000	394.63000	396.90000
LSTAT	4.98000	9.14000	4.03000	2.94000	5.33000
MEDV	24.00000	21.60000	34.70000	33.40000	36.20000

Ara es pot començar a intuir que els rangs formats pels valors de cada atribut són molt dispars. Tot i això, és molt difícil d'entendre-ho mirant simplement números i 5 registres. Ajudant-nos d'un diagrama de caixes es veu més clarament la distribució dels valors de cada atribut.



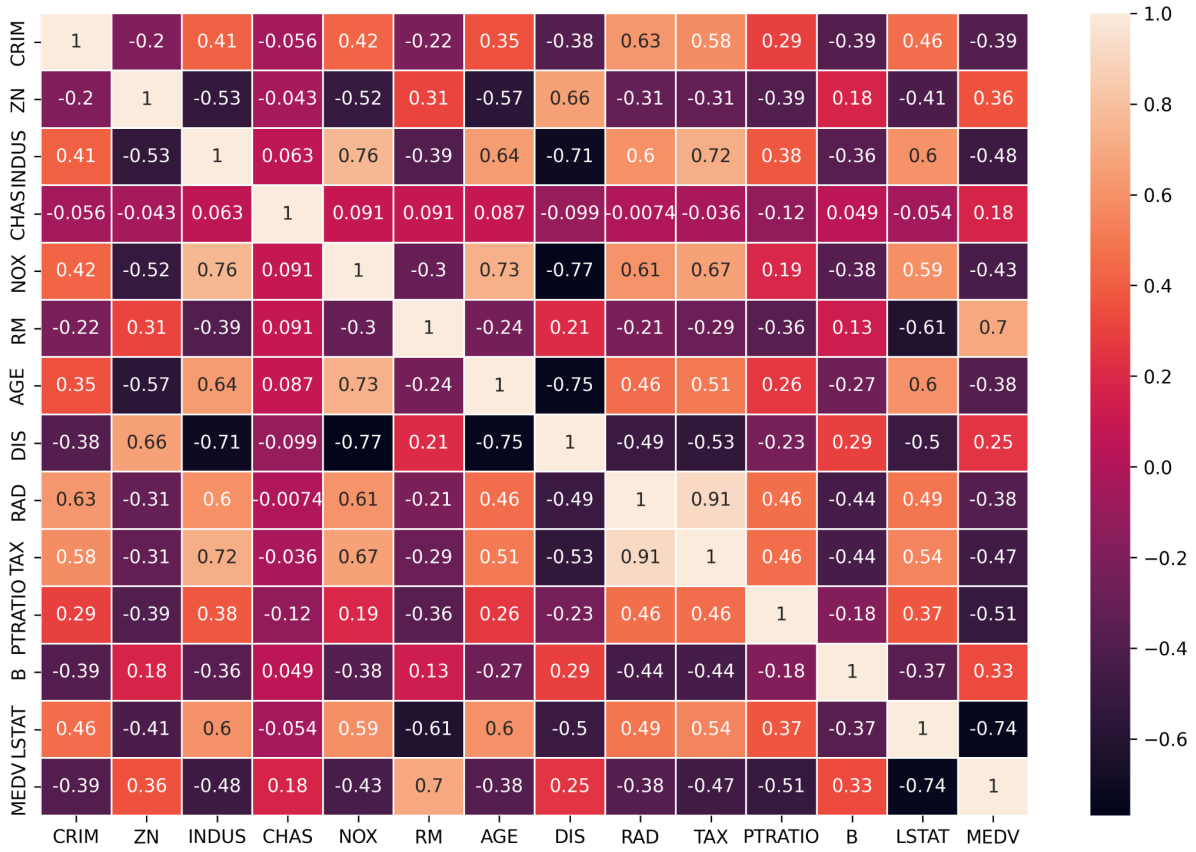
L'atribut TAX, per exemple, té valors molt més elevats que la resta. Passa quelcom similar, però menys significatiu, amb B i AGE. Més endavant es veurà que aquest fet pot resultar en un problema i com abordar-lo.

També es pot utilitzar un histograma per mostrar la distribució dels atributs. Ara es veu de forma clara el tipus de distribució de cada un. Els atributs RM i MEDV tenen una distribució Gaussiana.



La correlació entre dos atributs ens apropa encara més al nostre objectiu ja que ens permet veure quins atributs estan relacionats entre ells.

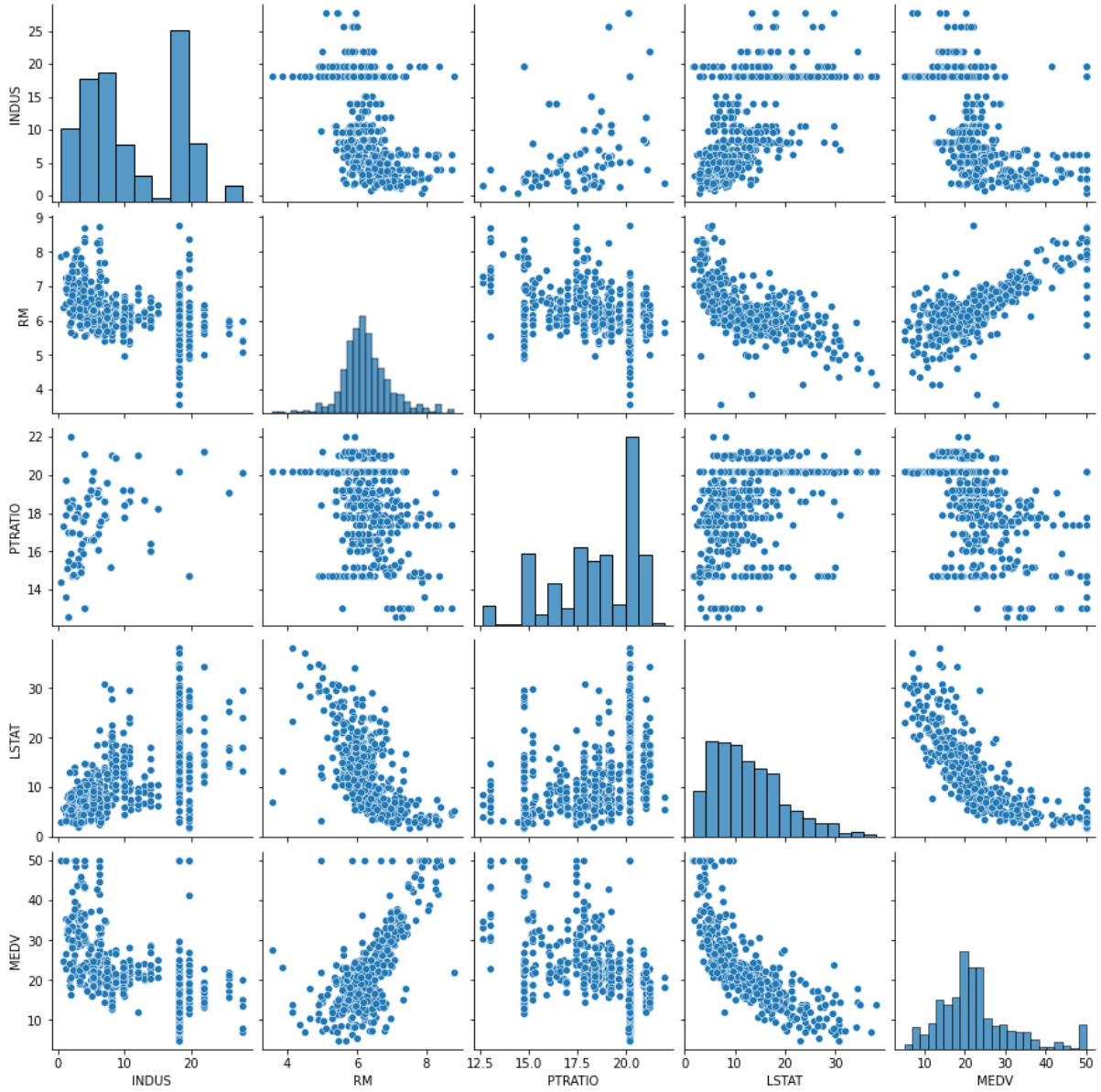
A la taula inferior es veu la correlació entre dos atributs. Els valors positius signifiquen correlació directa, mentre que els negatius signifiquen correlació inversa.



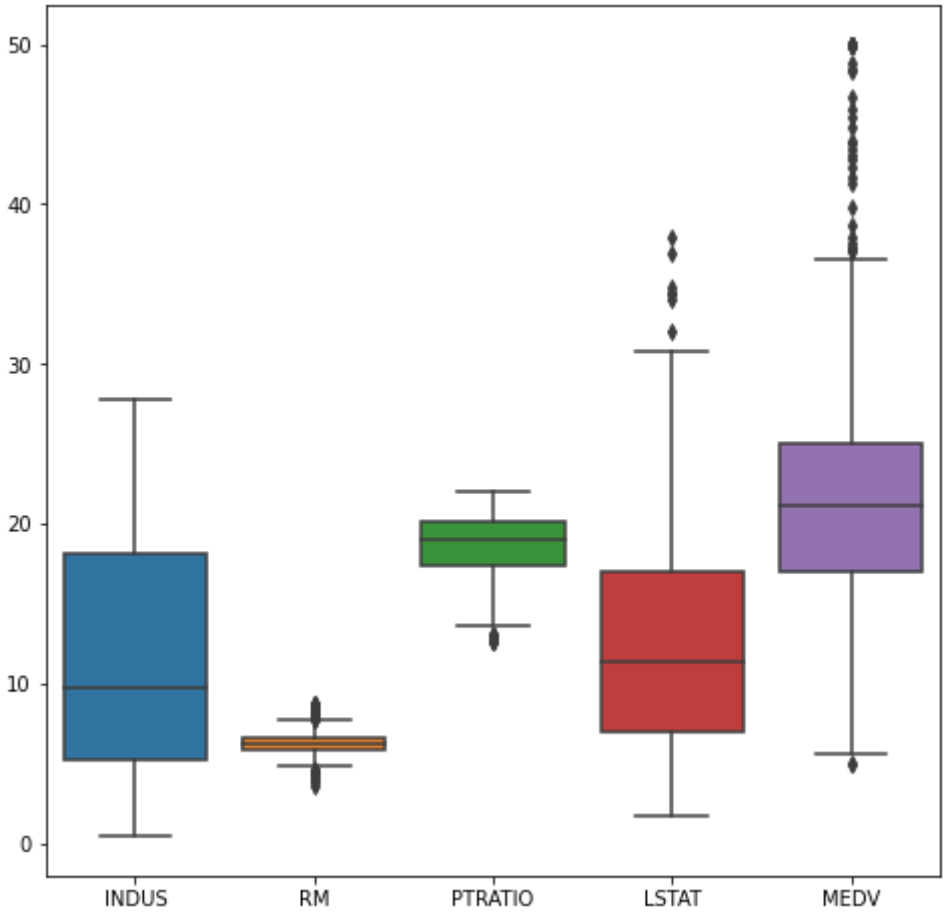
La taula permet veure quines característiques tenen correlació amb el preu de la casa. Una bona correlació es considera quan el valor absolut és igual o superior a 0.8. El nostre cas no és gaire favorable, doncs no hi ha cap característica que compleixi la condició.

Sí hi ha, però, LSTAT i RMv per sobre del 0.7 i INDUS, TAX i PTRATIO pròxims a 0.5. Aquests atributs seran els més importants per fer les prediccions, però amb la correlació que tenen ja podem fer-nos una idea que el resultat final no serà precís. Es provaran, si més no, diferents tècniques de tractament i priorització de dades per arribar al millor resultat.

A continuació es mostra un seguit de diagrames amb la distribució de cada característica i la correlació entre elles. Per Facilitar la comprensió només s'han utilitzat característiques amb una correlació superior a 0.5 a MEDV. Es pot apreciar molt bé que les característiques "LSTAT" i “RM “ tenen molta més correlació amb "MEDV" que la resta.



Ara que ja sabem quines característiques, a priori, són les més interessants, aprofitem per tornar a mostrar el diagrama de caixes, ara només amb els que tenen més correlació.



Si bé els valors de cada característica segueixen tenint diferents rangs, ja no hi ha l'abismal diferència que hi havia amb altres característiques ara considerades no correlatives.

Regressió Lineal

Primeres regressions

Per fer un aprenentatge es necessita dividir les dades en dos conjunts, Train i Test.
El Train és l'utilitzat per ensenyar al model i el Test és l'utilitzat per comprovar el seu funcionament.
És necessari fer dos conjunts per tal de fer proves amb dades que el model no hagi vist mai. Si s'utilitza un sol conjunt, el model ja l'ha vist durant l'aprenentatge i fa prediccions molt més bones.

Un cop es té una idea de les dades amb les que es tracta es pot començar a generar models de prediccions.

Es comença fent un Regressor Lineal per cada característica de la que es disposa i s'obtenen els següents resultats:

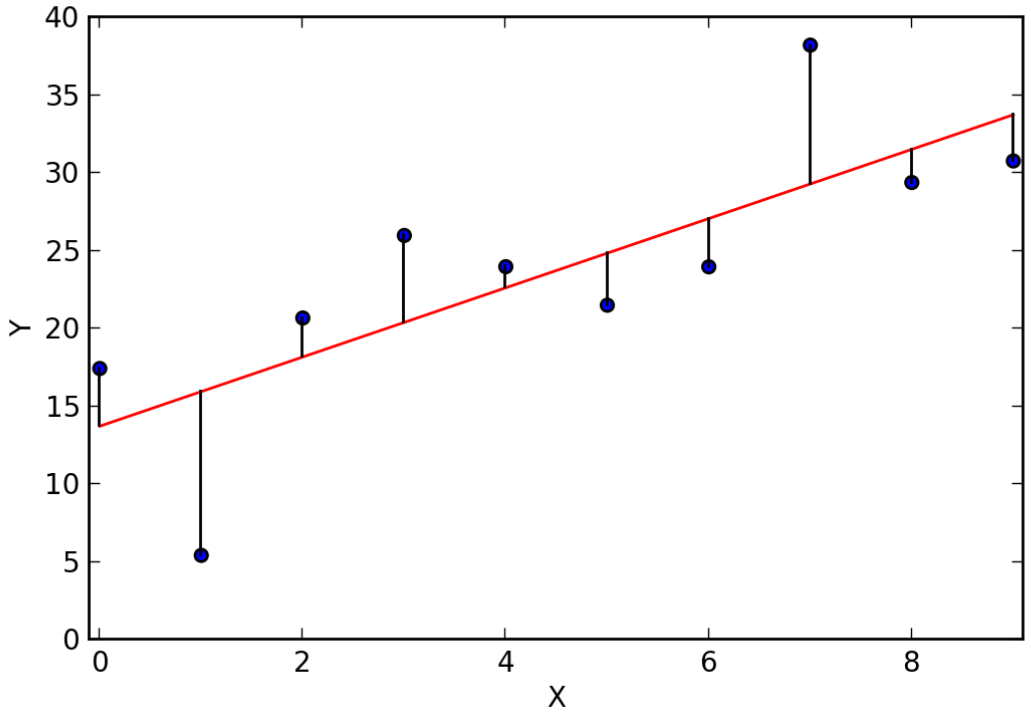
```
 CRIM
Coefficients: [-0.41097733]
Mean Squared Error:  58.96785248013031

 ZN
Coefficients: [0.13601374]
Mean Squared Error:  54.30823887705341

 INDUS
Coefficients: [-0.66699707]
Mean Squared Error:  52.98721928412714

 CHAS
Coefficients: [8.14953335]
Mean Squared Error:  72.86023296145912

 NOX
Coefficients: [-33.7843635]
Mean Squared Error:  55.74707753668165

 RM
Coefficients: [9.55621428]
Mean Squared Error:  48.18702933783806

 AGE
Coefficients: [-0.11963822]
Mean Squared Error:  54.70995150687366

 DIS
Coefficients: [1.04860501]
Mean Squared Error:  61.59074757916875

 RAD
Coefficients: [-0.4420297]
Mean Squared Error:  66.83511282112357

 TAX
Coefficients: [-0.0272483]
Mean Squared Error:  60.73697741742185

 PTRATIO
Coefficients: [-2.21271014]
Mean Squared Error:  51.80542865366026

 B
Coefficients: [0.03387417]
Mean Squared Error:  61.435113407538395

 LSTAT
Coefficients: [-0.9456965]
Mean Squared Error:  29.145028348828678
```

Per saber si un resultat és bo o dolent cal fixar-se en el Mean Squared Error. Per entendre'l es pot utilitzar el gràfic inferior on es veu un seguit de punts que indiquen el valor predit. La diagonal vermella és el valor real de cada punt. Les verticals negres mostren la distància a la que s'ha quedat la predicció.

L'error pot ser negatiu o positiu, així doncs s'ha de fer un sumatori de l'error quadràtic de cada punt i dividir el resultat pel nombre total de valors predits.

Per entendre el resultat només s'ha de fer l'arrel quadrada del MSE i es veurà, de mitjana, quin error de predicció dona.

$$\underbrace{\text{MSE} = \frac{1}{n}}_{\text{test set}} \sum_{i=1}^{n} \underbrace{(y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i)^2}_{\text{actual value}}$$



Com era previsible, com més correlació té una característica, el MSE és més petit.

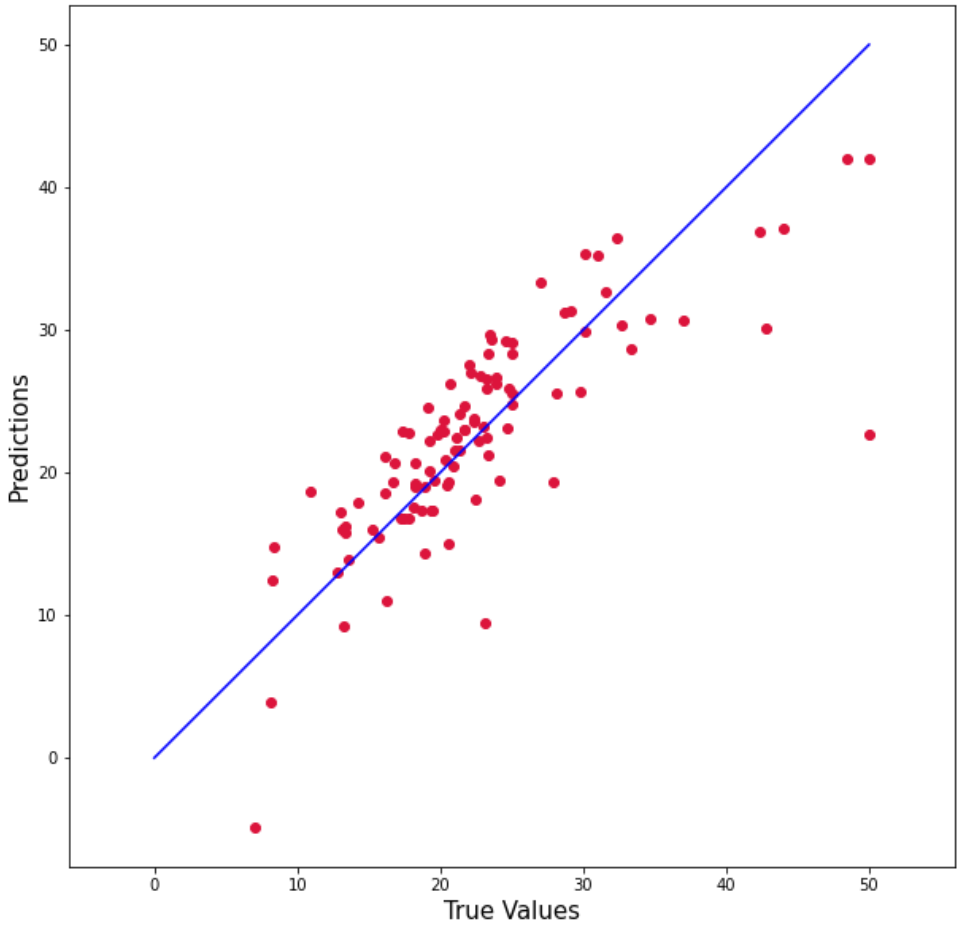
La millor predicció s'ha fet amb la característica "LSTAT", obtenint un MSE = 29.14. Calculant l'arrel quadrada surt un resultat de 5.4.

Per tant, com que l'atribut MEDV està en milers, de moment podem predir el preu d'una casa amb un error mitjà de +-5,400 $.

Regressió Lineal amb totes les característiques

Fins ara només s'ha utilitzat uns sola característica per fer la regressió lineal. Una altra opció que acostuma a donar millors resultats és utilitzar totes les característiques.

El gràfic mostrat a continuació permet veure l'error de cada una de les prediccions. L'eix X és el valor real de la casa, mentre que l'eix Y és el valor que s'ha predit. La línia horitzontal ens permet veure la distància vertical (diferència) que hi ha entre el valor predit i el valor real. Es pot utilitzar per calcular de forma manual l'error quadràtic. Durant la resta de l'apartat de regressió lineal s'utilitzarà aquest tipus de gràfic.

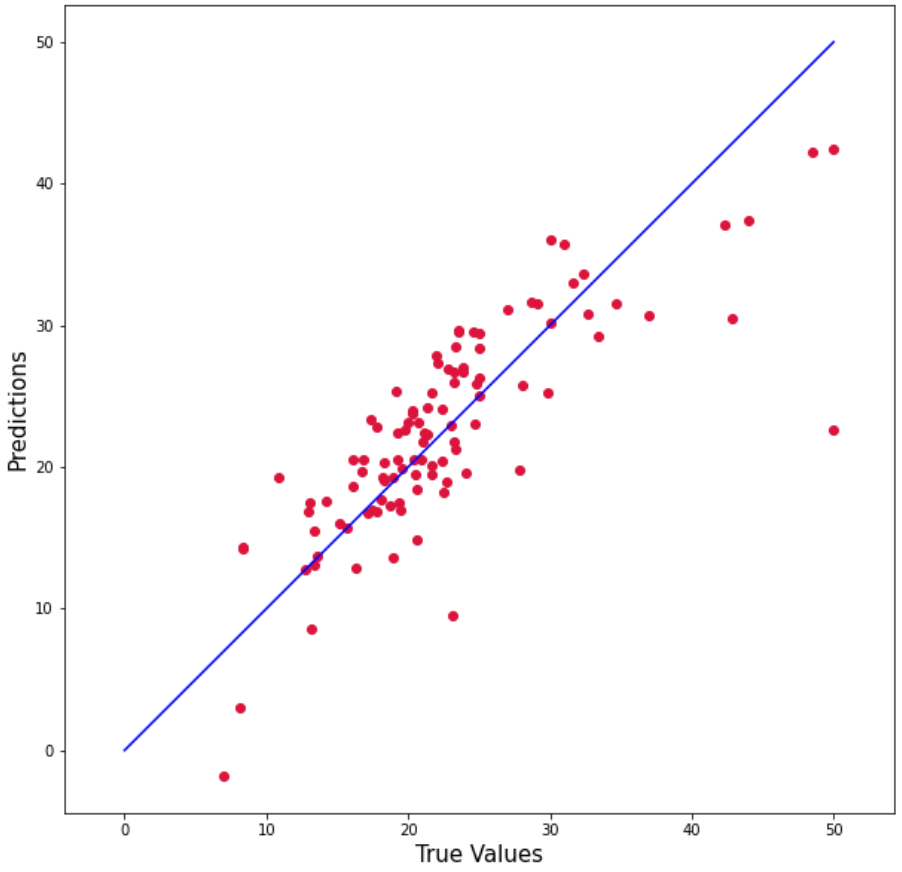



```
Coefficients: [-1.21942588e-01  3.72364150e-02  2.72655529e-02  3.10158235e+00
 -1.79082913e+01  4.18063065e+00 -5.80131969e-03 -1.58082749e+00
  3.02006011e-01 -1.30894107e-02 -1.02318767e+00  1.02695918e-02
 -4.78095078e-01]
Intercept: 35.84221860695473
Training MSE:  21.32652088949429
Test MSE:  25.109737423651612
Coefficient of determination:  0.6322347535411479
```

Amb aquest model s'obté un MSE de 25.10, que es tradueix a un error mitjà de +-5000 $.

Regressió lineal amb característiques importants

A continuació es genera un nou model, però aquest cop s'eliminen característiques amb poca correlació i que podrien estar generant soroll. Es tracta de CRIM, CHAS i AGE.



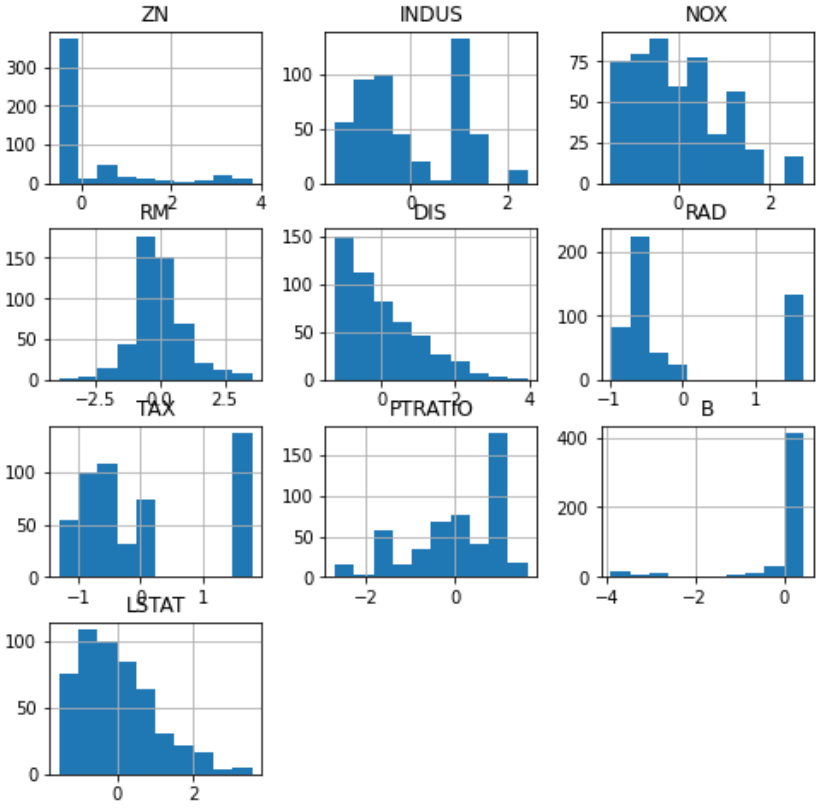

```
Coefficients: [ 3.30588385e-02  6.09378715e-02 -1.74076384e+01  4.24726200e+00
 -1.49446602e+00  2.65856382e-01 -1.47791577e-02 -1.07608529e+00
  1.17241509e-02 -5.22075133e-01]
Intercept: 35.90532442511293
Training MSE:  22.645913681970935
Test MSE:  24.545222759917458
Coefficient of determination:  0.6405028158842561
```

L'error quadràtic fa una petita millora fins a 24.5. Significa que ara l'error per preveure el preu d'una casa és de +-4900 $.

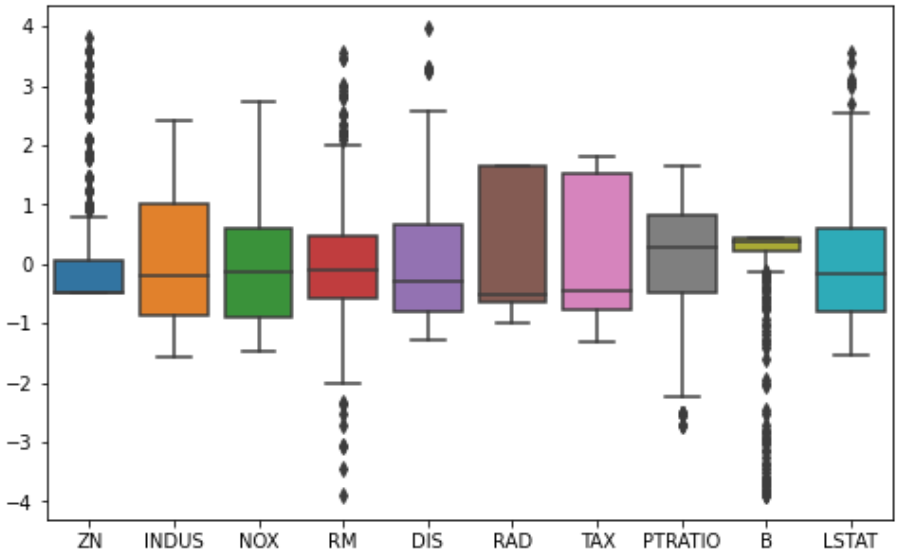
Estandaritzant valors

Fins ara s'ha estat fent l'entrenament del model amb valors sense tractar. Hi ha característiques amb rangs molt amples que segurament estan influint en els pesos de cada atribut durant la regressió. Aplicant una estandardització aconseguim rangs similars per totes les característiques.

El primer gràfic permet veure el diagrama de freqüències de cada característica. Es pot apreciar com, tot i tenir la mateixa distribució, els valors de les dades estan tots entre rangs similars.

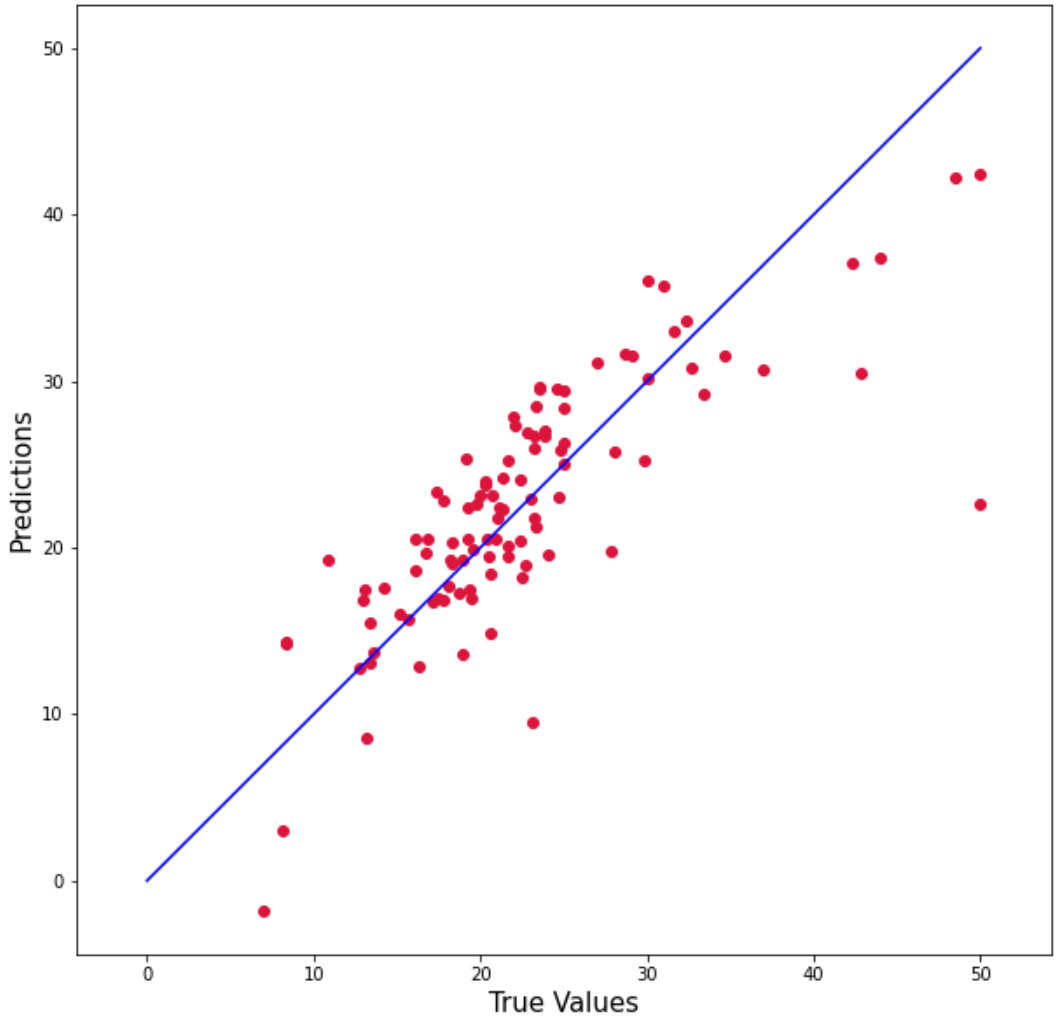


La segona imatge mostra de forma més gràfica els rangs estandarditzats entre les diferents característiques.



Regressió Lineal amb característiques estandarditzades

Es fa un nou Regressor Lineal però s'obté, contra tot pronòstic, el mateix error quadràtic.




```
Coefficients: [ 0.77025096  0.417642   -2.01516245  2.98124882 -3.1438011   2.31259191
 -2.4883741  -2.32736285  1.06929658 -3.72448485]
Intercept: 22.546131990002685

Training MSE:  22.645913681970935
Mean Squared Error:  24.545222759917582
Coefficient of determination:  0.6405028158842542
```

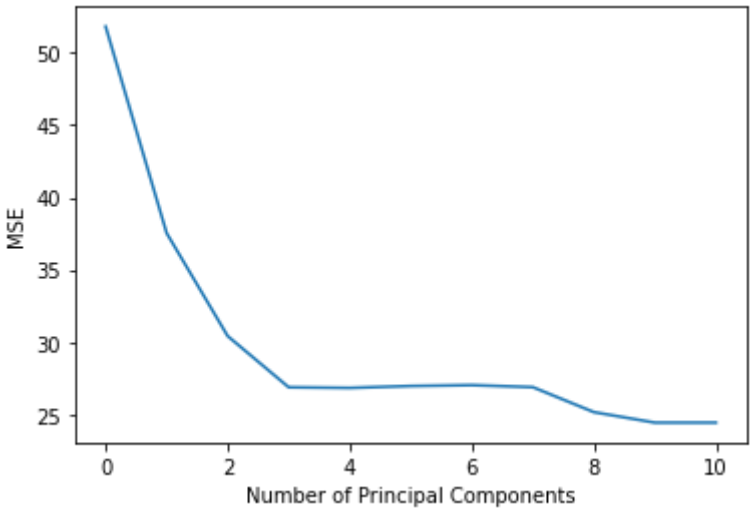
El resultat no canvia perquè les característiques que tenen rangs més dispars són les que tenen menys correlació amb "MEDV".

Buscant nombre de components màxim

Un mètode per reduir les dimensions de les característiques és el Principal Component Analysis.

Fent servir el PCA es busca quin és el número de components generats que dona el MSE més petit. Per assegurar-se que ho fa bé i no és pas fruit de l'atzar durant el split entre el test i el train, per cada component es fa un cross-validation amb 10 splits i 10 repeticions.

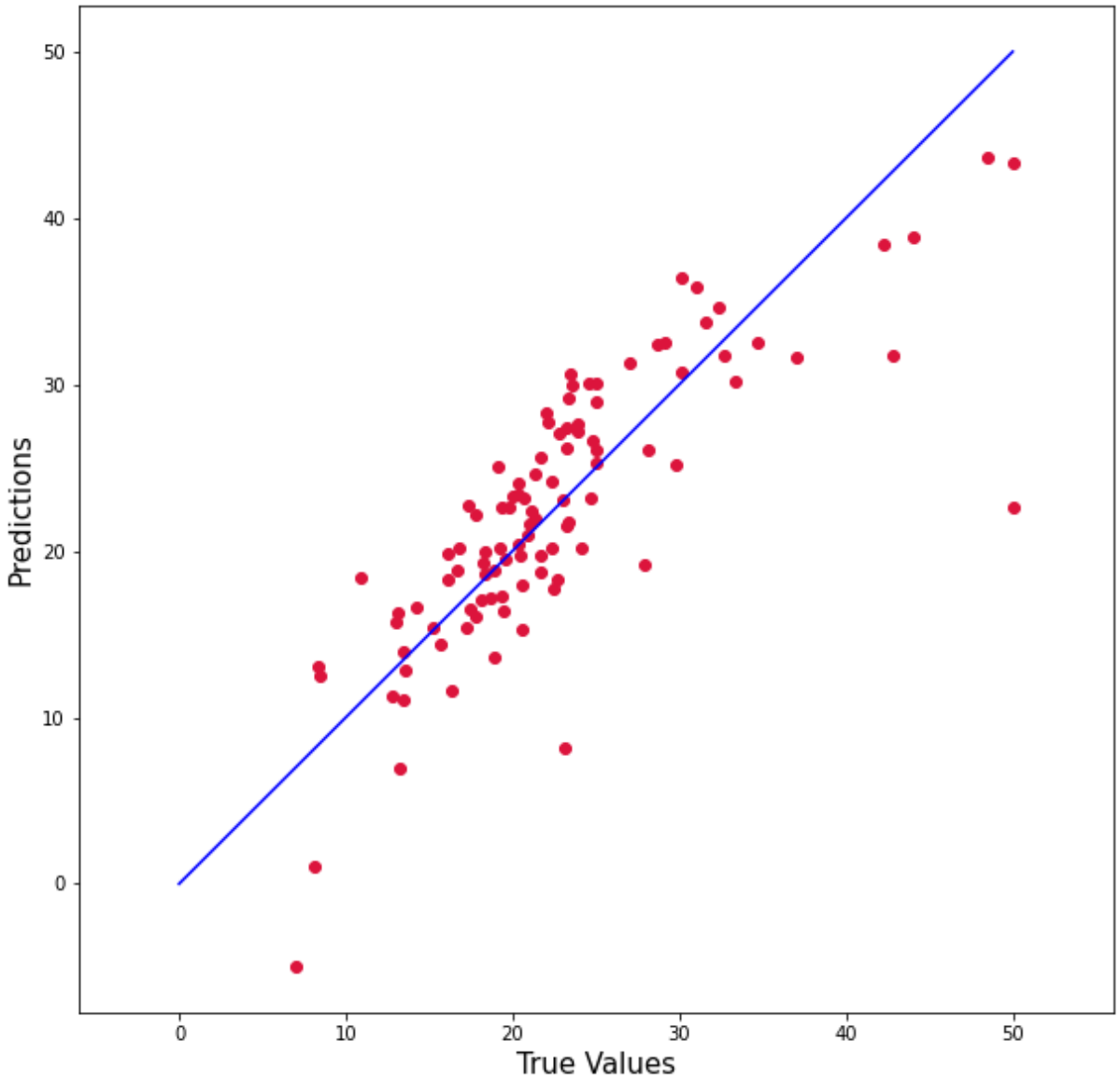
S'obté la següent gràfica:



En aquest cas el PCA no és capaç de generar components que el seu nombre faci millorar el MSE de forma substancial. El nombre de components que genera un MSE millor és 10.

Aquest fet passa perquè el número característiques tampoc és tant elevat i no hi ha gaire redundància al dataset. El PCA, per tant, no és capaç de reduir la dimensionalitat sense reduir també la variança.

Fent un nou model de Regressió Lineal on la X és el conjunt de components generats pel PCA quan n_components = 10 surt el següent resultat.



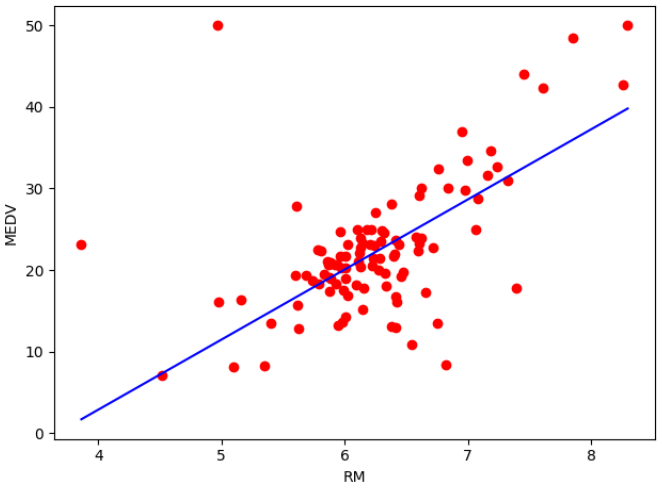

```
PCA number of components:  10
Coefficients: [ 2.57808414  4.0300456  -2.15777738 -2.07589746 -0.99909885  0.0420904
  0.39756697  0.9341842  -3.28026921 -3.50050507]
Intercept: 22.521039603960396
Test MSE:  25.28923530881304
Coefficient of determination:  0.6296057700969381
```

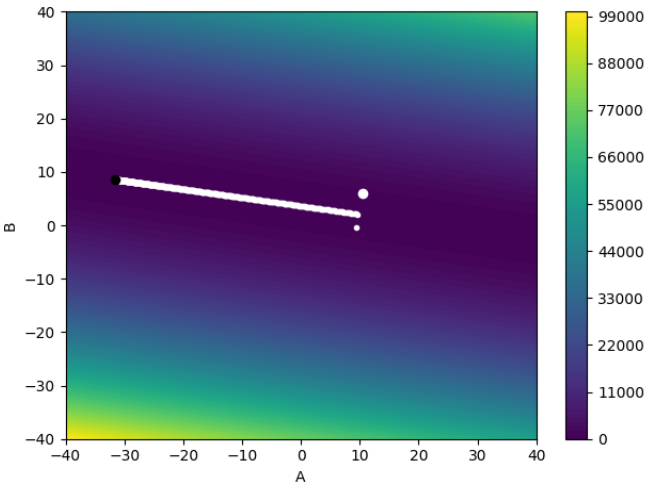
El MSE s'apropa al resultat obtingut anteriorment, però no és capaç d'igualar-lo.

Descens de gradient

Implementació Pròpia

Per fer el descens de gradient hem utilitzat la fórmula del mse com a funció a minimitzar ja que busquem tenir el menor error possible.
Per fer-ho hem hagut de crear una funció que el calculés a la que direm cost (o *func()* al codi). També hem de definir una funció de gradient, en aquest cas utilitzem les derivades parcials per saber quin sera el pròxim pas.





cost on test: 46.454083241013194 th: [-31.5709899 8.60505868]

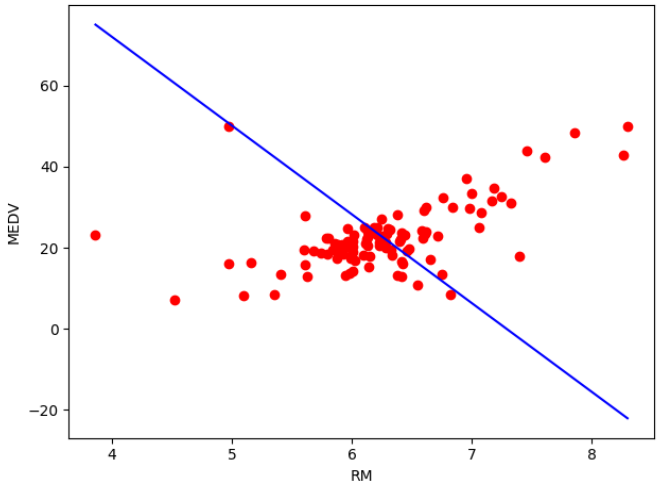
Aquests són els resultats fent el descens de gradient amb MEDV i RM.
Primer creem un punt a unes coordenades aleatòries. Després és quan comencem a calcular les següents posicions del punt, primer afegint el valor de alpha a Theta per seguidament calcular el gradient amb la derivada parcial.

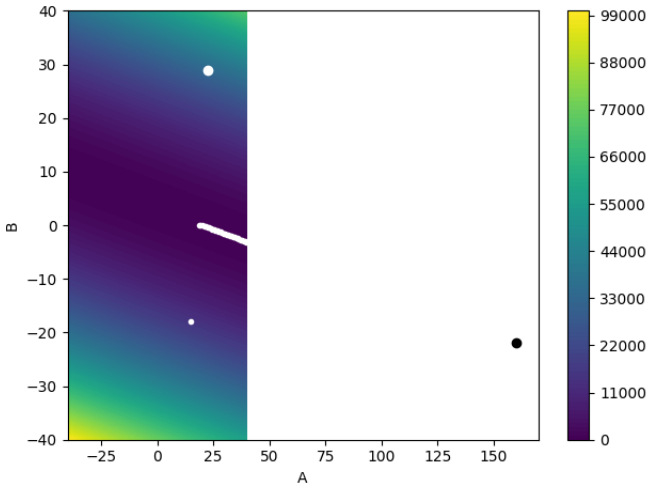
Els resultats mostrats anteriorment han estat fets amb els següents paràmetres:

- alpha = 0.001
- lr = 0.02
- nRep = 5000

alpha: determina la “mida” de les pases que pot fer el punt.
lr (learning rate): es multiplica pel gradient, que determinarà la següent passa.
nRep: es el nombre de repeticions que es fan.

Una *alpha* massa gran o un *lr* massa gran farà que el cost aumenti i com a conseqüència no trobarà un mínim local.

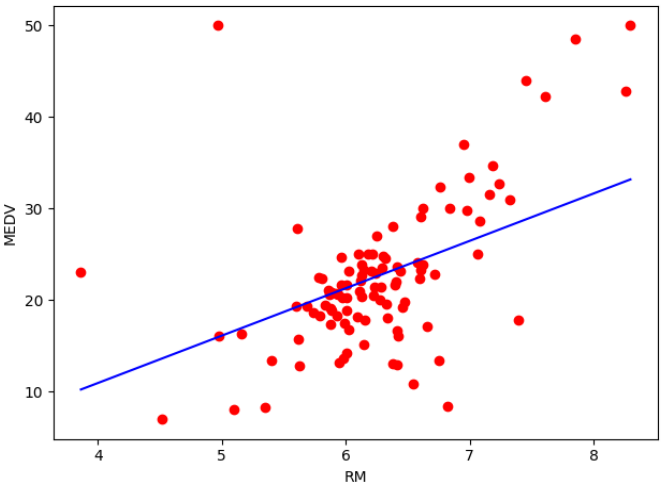


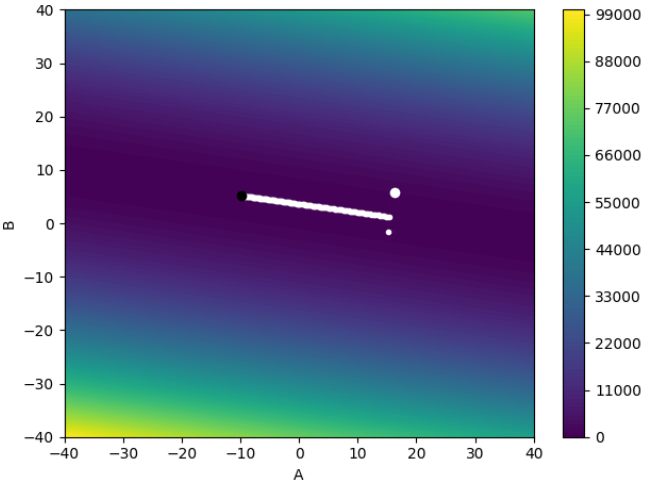


- cost on test: 426.79823623644165 th: [159.85860017 -21.9361682]
- alpha=1.0
- lr=0.02
- nRep=5000

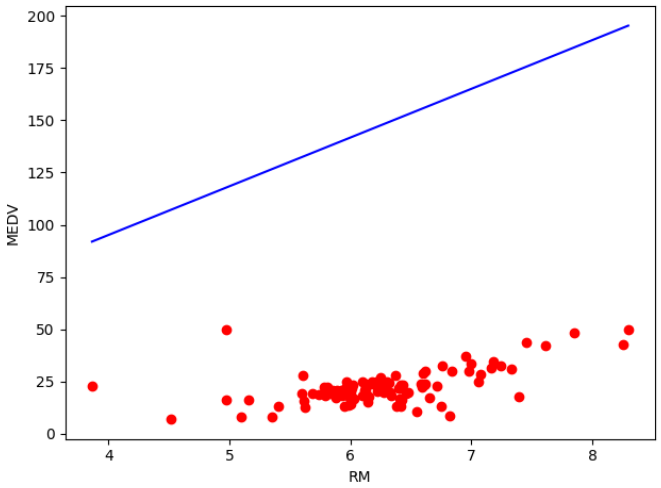
Com podem veure, la recta que hauria de predir el valor de la casa en funció del nombre d'habitacions no es la més òptima tot i tenir una bona correlació.

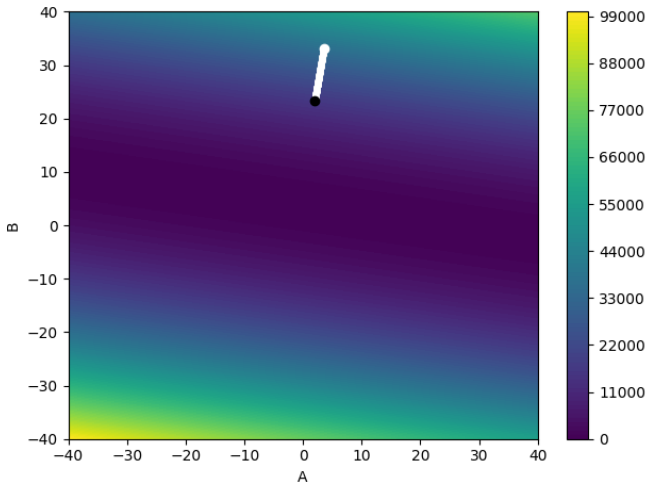
Però utilitzar una *alpha* massa petita o una *lr* massa petita pot fer que mai es trobi un mínim local ja que es detindrà més fàcilment a punts no gaire òptims. Per altra banda, necessita moltes repeticions.





- cost on test: 47.08711780611418 th: [-9.74095864 5.17572133]
- alpha=0.0000000000001
- lr=0.02
- nRep=5000

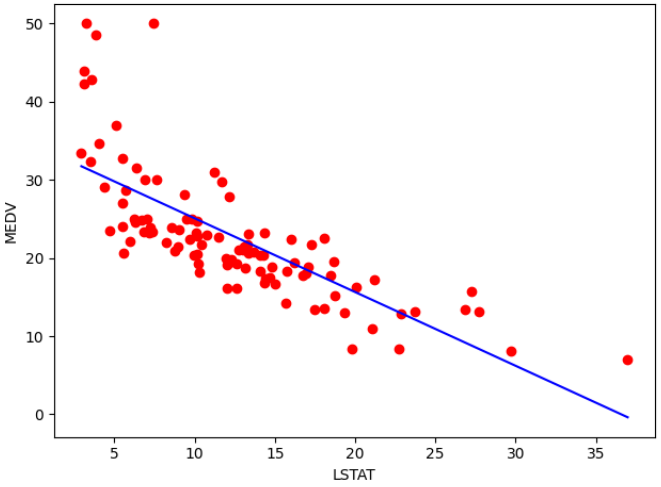


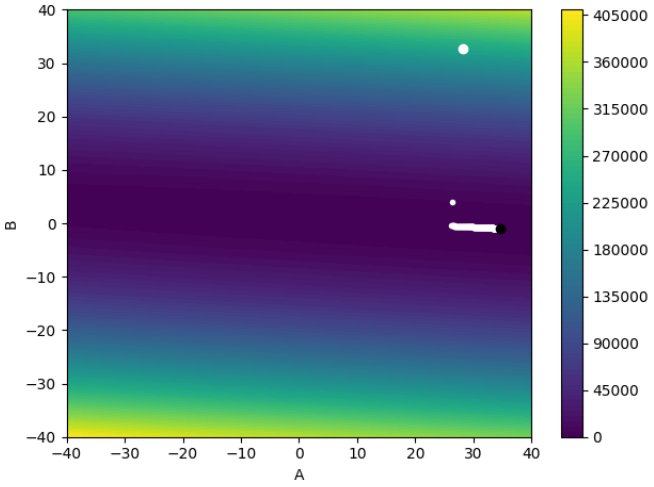


- cost on test: 15840.288427266525 th: [1.99189042 23.29296554]
- alpha=0.001
- lr=0.000001
- nRep=5000

Al gràfic de la dreta podem veure com s'ha quedat a mig cami de una solució mínimament viable.
També podem observar que la variable lr té un major impacte en el resultat.

Reduir el nombre de repeticions empitjoraria els resultats en general.



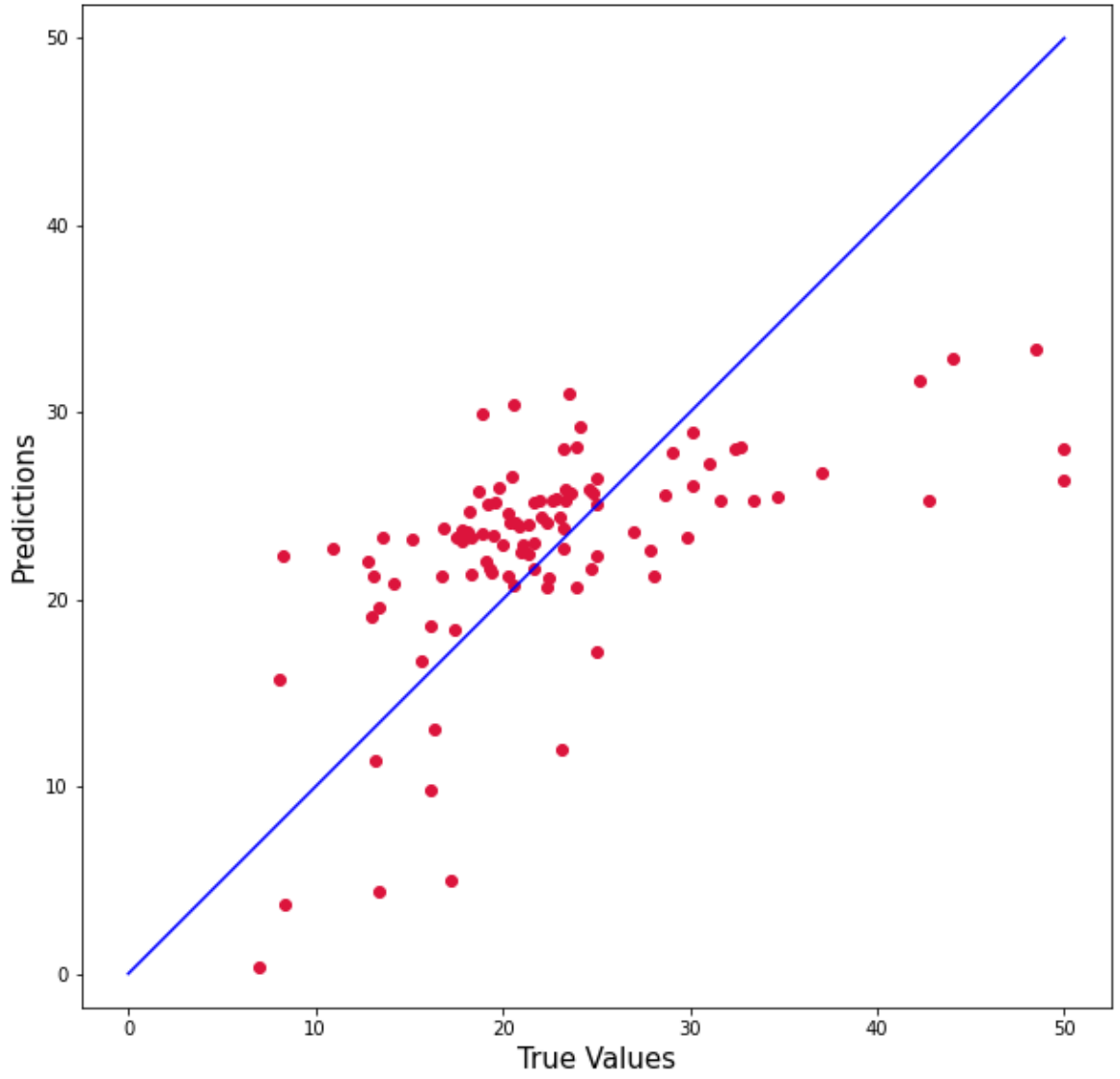


- cost on test: 29.1401113235097 th: [34.51989343 -0.94399742]
- alpha=0.001
- lr=0.002
- nRep=5000

Implementació amb SGDRegressor

També s'ha volgut veure quin resultat s'obtindria utilitzant paquets de Sklearn. Per fer-ho s'ha utilitzat el SGDRegressor.

Per mostrar els resultats tornem a utilitzar el tipus de gràfic anterior que mostra la distància entre el punt predit i el seu valor original.




```
Mean Squared Error:  46.27885423390113
Coefficient of determination:  0.3221850971196326
```

Webgrafia

- https://www.youtube.com/watch?v=HMOI_lkzW08
- https://www.baeldung.com/cs/normalization-vs-standardization
- https://stats.stackexchange.com/questions/10289/whats-the-difference-between-normalization-and-standardization
- https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60
- https://stats.stackexchange.com/questions/451619/what-does-it-mean-when-pca-does-not-produce-a-reduction-in-dimensionality
- https://scikit-learn.org/stable/modules/sgd.html
- https://www.datatechnotes.com/2020/09/regression-example-with-sgdregressor-in-python.html
- https://scikit-learn.org/stable/
- https://towardsdatascience.com/why-gradient-descent-and-normal-equation-are-bad-for-linear-regression-928f8b32fa4f
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html

Github Link

- https://github.com/marti1999/APC-GPA203-0930-PLAB1