Studio di Analisi dei Dati

Francesco Mergiotti

25 settembre 2017

Contents

Introduzione al Dataset

Il dataset studiato è una raccolta di 506 osservazioni riguardanti dati su proporzioni della popolazione della città di Boston. Ogni osservazione si riferisce a zone specifiche e riporta informazioni sullo status generale degli abitanti e sulle condizioni della zona stessa.

Le features che compongono il dataset sono 13 e ora le vediamo in dettaglio:

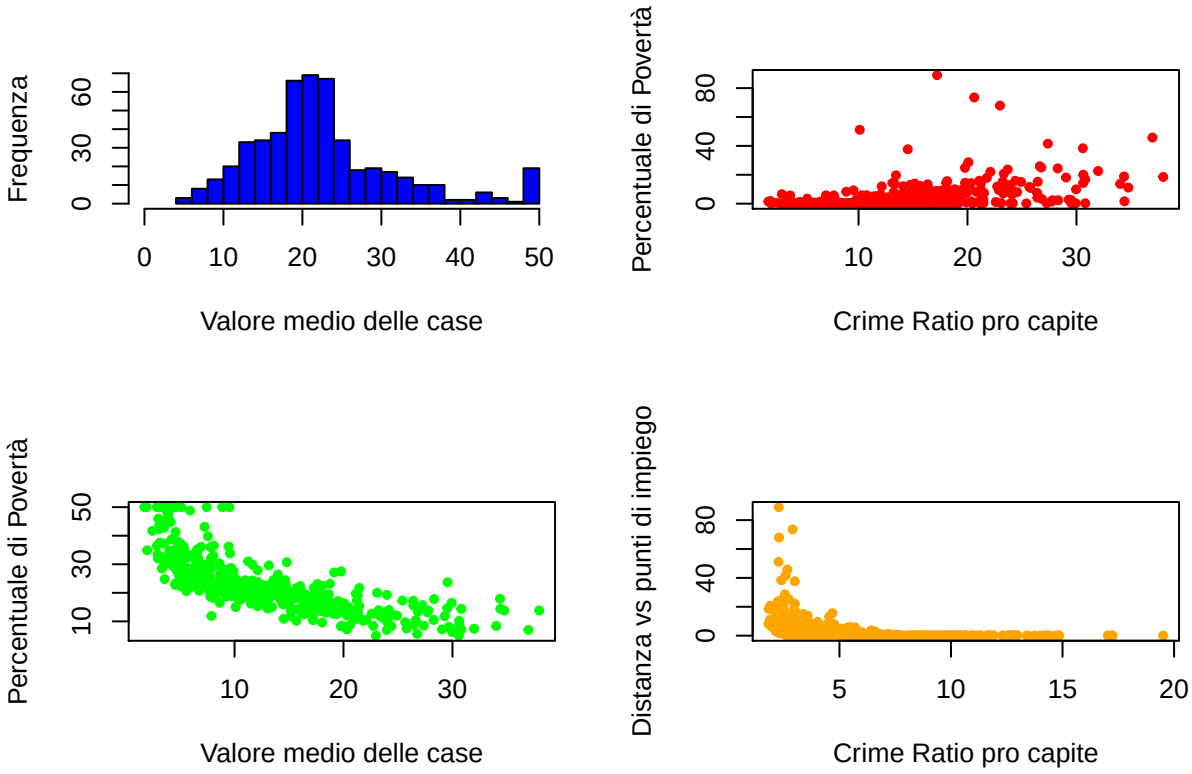
- CRIM - percentuale di crimine pro capite
- ZN - proporzione della zona residenziale (25.000 sq.ft)
- INDUS - proporzione di commercianti all'ingrosso
- CHAS - zona limitata dal fiume Charles (1 se limitata, 0 altrimenti)
- NOX - concentrazione di monossido d'azoto
- RM - media di stanze per abitazione
- AGE - percentuale di case costrutite prima del 1940 occupate dai proprietari
- DIS - distanza media dai 5 centri di impiego di Boston
- RAD - indice di accessibilità alle autostrade
- TAX - Valore totale della tassa sulla proprietà x$10,000
- PTRATIO - percentuale studente-insegnante
- BLACK - 1000(Bk - 0.63)^2 dove Bk è la percentuale di neri in città
- LSTAT - percentuale di popolazione con un basso status sociale
- MEDV - Valore medio delle case occupate x$1,000

Pulizia Dati

Si è eliminata la colonna nox, che rappresenta la concentrazione di monossido d'azoto, che ai fini dello studio del suddetto dataset è poco rilevante. Abbiamo inoltre modificato la colonna zn relativa alla proporzione della zona residenziale ogni 25.000 piedi quadrati, convertita in metri quadrati. La colonna dis è stata riportata da miglia a km.

Analisi iniziale dei dati

L'analisi sommaria iniziale è stata svolta considerando delle features che, a nostro parere, sono sembrate più significative di altre. Di seguito riportiamo qualche plot.



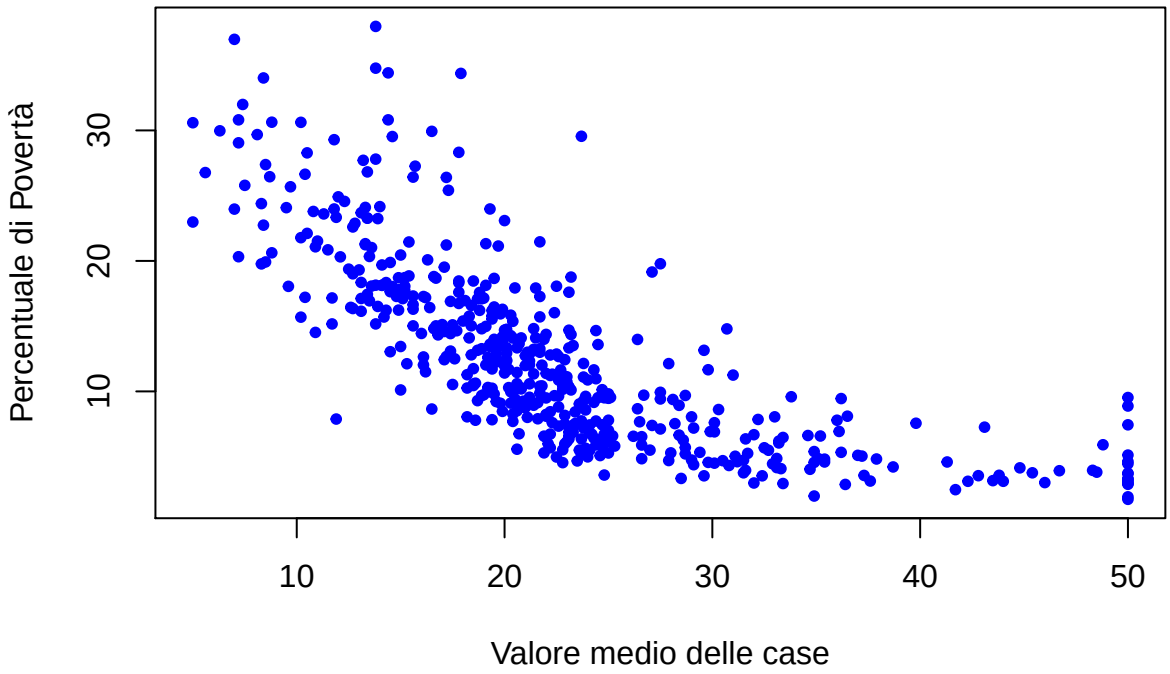
Prime osservazioni

Le prime osservazione sono state sviluppate considerando le feature più significative:

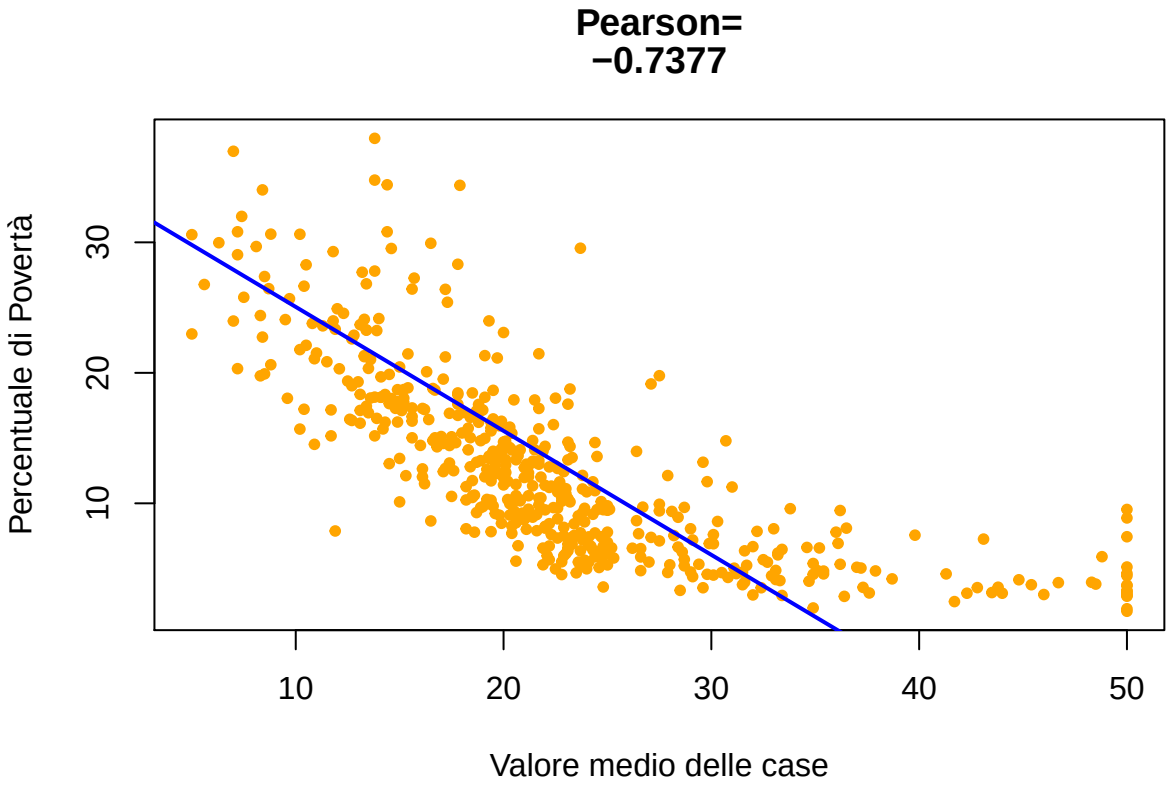
- Valore medio delle case di Boston
- Percentuale dello status sociale basso
- Ratio del crimine pro capite

Correlazione tra valore medio delle case e la percentuale di povertà

Una prima correlazione, che è saltata all'occhio analizzando il dataset, è quella tra il valore medio delle case di Boston e lo status sociale della popolazione. Entrambi, infatti, sono sufficientemente correlati, come si può notare dal primo plot sperimentale:



Abbiamo eseguito la prova del nove utilizzando i coefficienti di Pearson e Spearman, il secondo è stato utilizzato per rafforzare la correlazione.



Il primo coefficiente da un valore di -0.74, in accordo con il primo fit stampato poco sopra.

```
cor(BB$ValoreCase,BB$Poverta, method="spearman")
```

```
## [1] -0.8529141
```

Spearman rafforza questa correlazione portando il valore a -0.85, dando un ottimo risultato. Da questa prima analisi possiamo dedurre che lo status sociale della popolazione è direttamente correlato con il valore medio delle case di Boston. Vale anche il viceversa.

Abbiamo scoperto che MEDV e LSTAT sono fortemente correlato, quindi possiamo prevedere la percentuale dello status sociale basso della popolazione in base al valore medio delle case.

Correlazione crimine e valore medio delle case

Abbiamo anche studiato la relazione tra il tasso di crimine e il valore medio delle case. Il risultato è stato deludente e poco convincente.

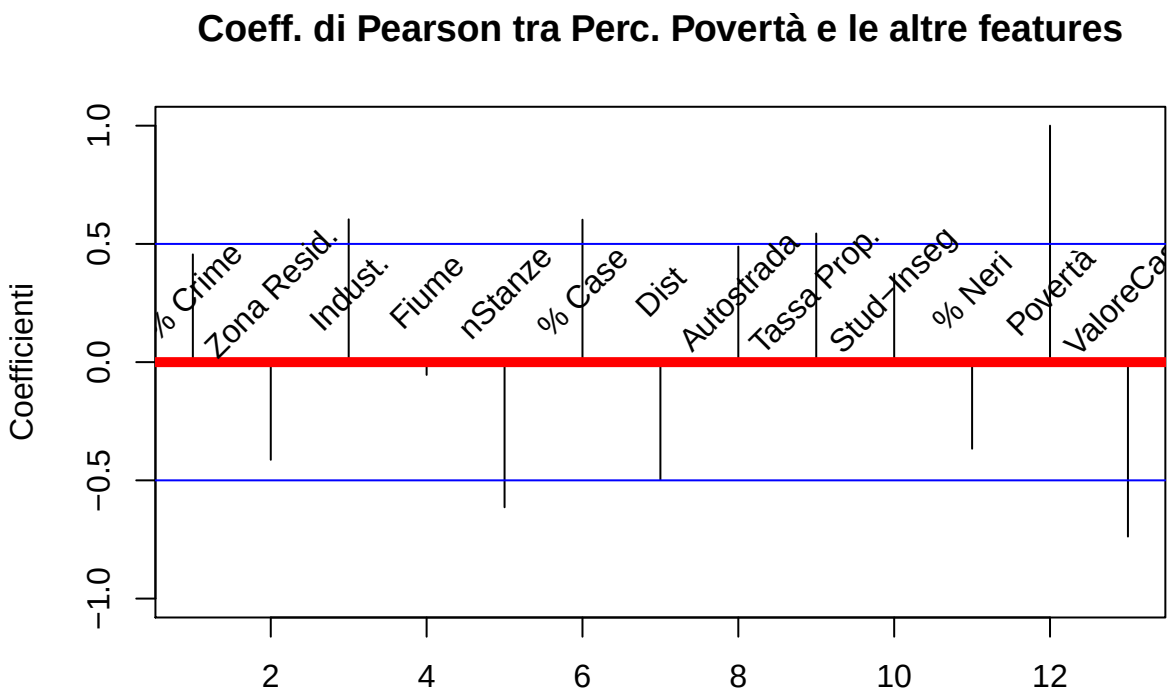
```
cor(BB$CrimRatio,BB$ValoreCase, method="pearson")
```

```
## [1] -0.3883046
```

Pressochè bassa la correlazione. Scartiamo questa opzione di verifica.

Studio dei regressori

Abbiamo studiato quali regressori potrebbero prevedere al meglio lo Status sociale basso della popolazione. Il confronto con le altre feature è riassunto in questo plot:



Da questa analisi possiamo considerare le features che superano la soglia arbitrariamente fissata da noi a -0.5 e 0.5. Nelle analisi future considereremo le seguenti componenti:

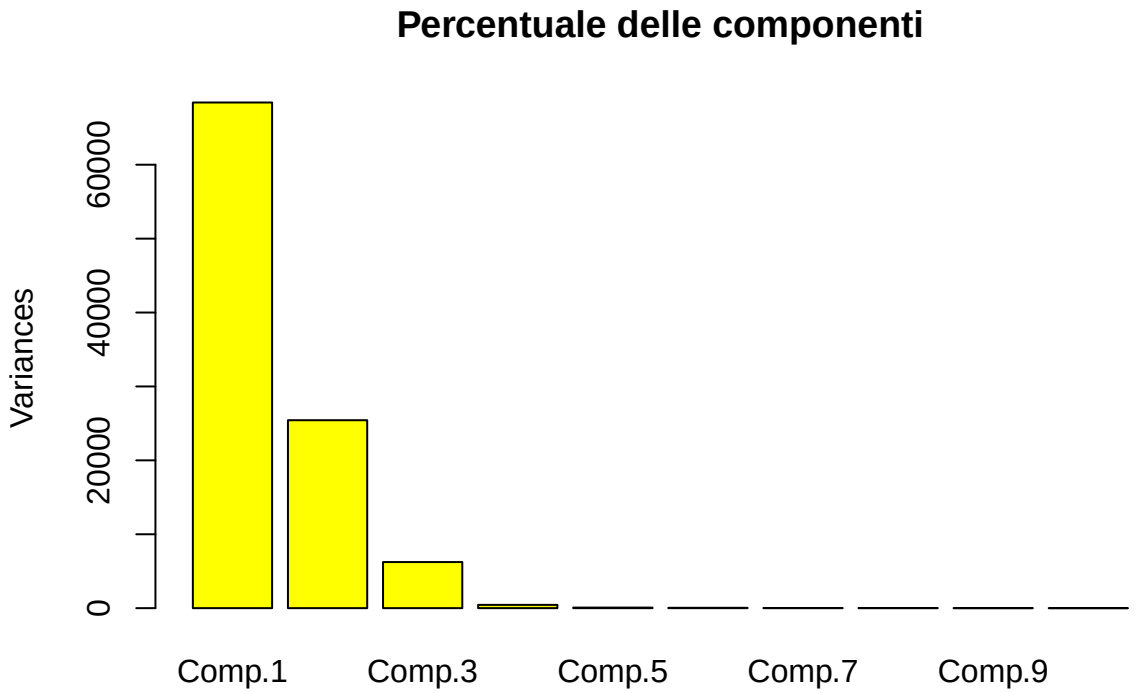
- Valore medio delle case = -0.7376
- N stanze per abitazione = 0.6138
- Industrie = 0.6037
- % case abitate dai proprietari = -0.6023
- Tassa sulla proprietà = 0.5439
- Distanza dai centri di impiego di Boston = -0.4969

Vogliamo prevedere la percentuale di persone che versano in un basso stato sociale utilizzando questi parametri. Possiamo quindi definire un nuovo campo come livello di povertà con questi 4 attributi:

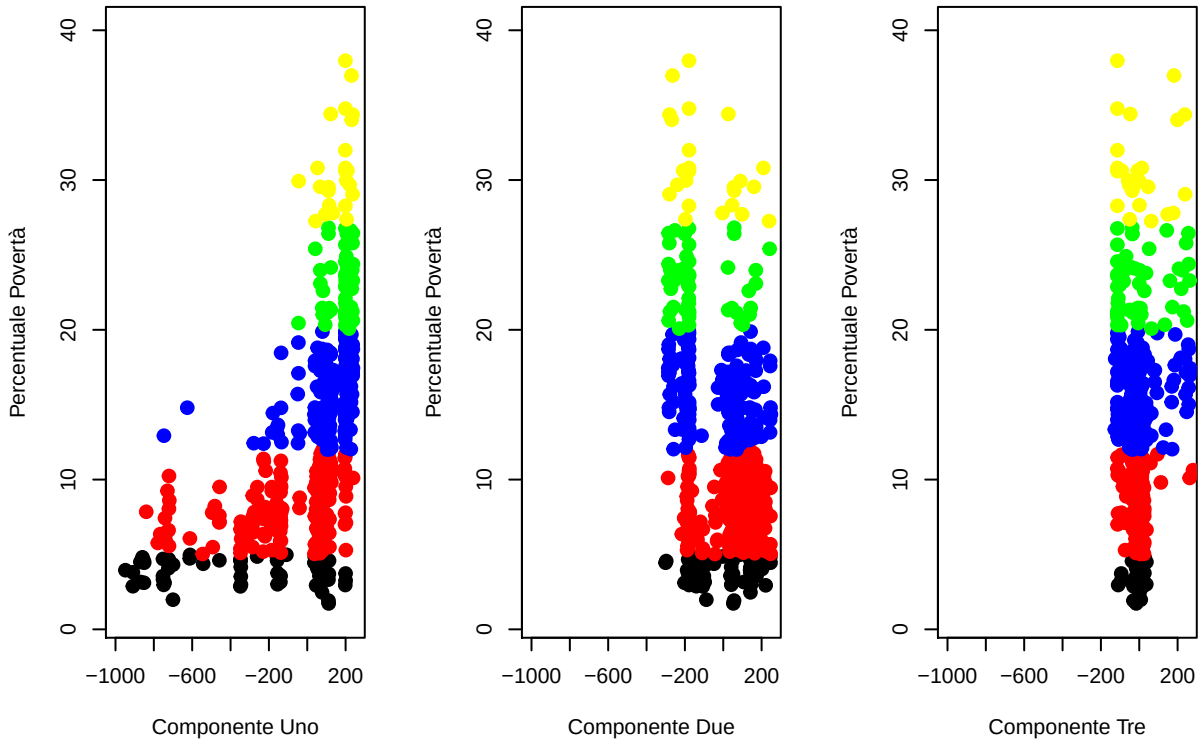
- 1% - 5% : povertà inesistente
- 6% - 12% : povertà bassa
- 13% - 20% : povertà discreta
- 21% - 27% : povertà accentuata
- 28% - 39% : povertà elevata

Compressione del dataset

Abbiamo notato con l'analisi precedente che per prevedere la percentuale della feature LSTAT, necessitiamo di 5 componenti su 12 totali. Possiamo dunque ridurre il nostro dataset e cercare di sfoltire la mole di dati su cui lavoriamo. Il metodo che abbiamo utilizzato è la Principal Component Analysis - PCA.



Con questo plot riusciamo a vedere che le prime 3 componenti riescono a spiegare ben il 99,38 % del dataset. Dunque procediamo a sfoltire il nostro malloppo di dati tenendo solo le prime 3. Qui di seguito riporto i grafici della percentuale di spiegazione



Osservazione

Avremo potuto tenere solo le prime due componenti e ricavare un risultato del 93%. Considerando che la mole di dati è bassa e con l'aggiunta di un'altra componente si arriva a spiegare il 99% del dataset, si è deciso di includere anche la terza componente, che funge in questo caso da cornice per il nostro quadro.

Classificatori

In questa fase dello studio di questo dataset, tentiamo di utilizzare vari metodi per spiegare e prevedere la feature LSTAT.

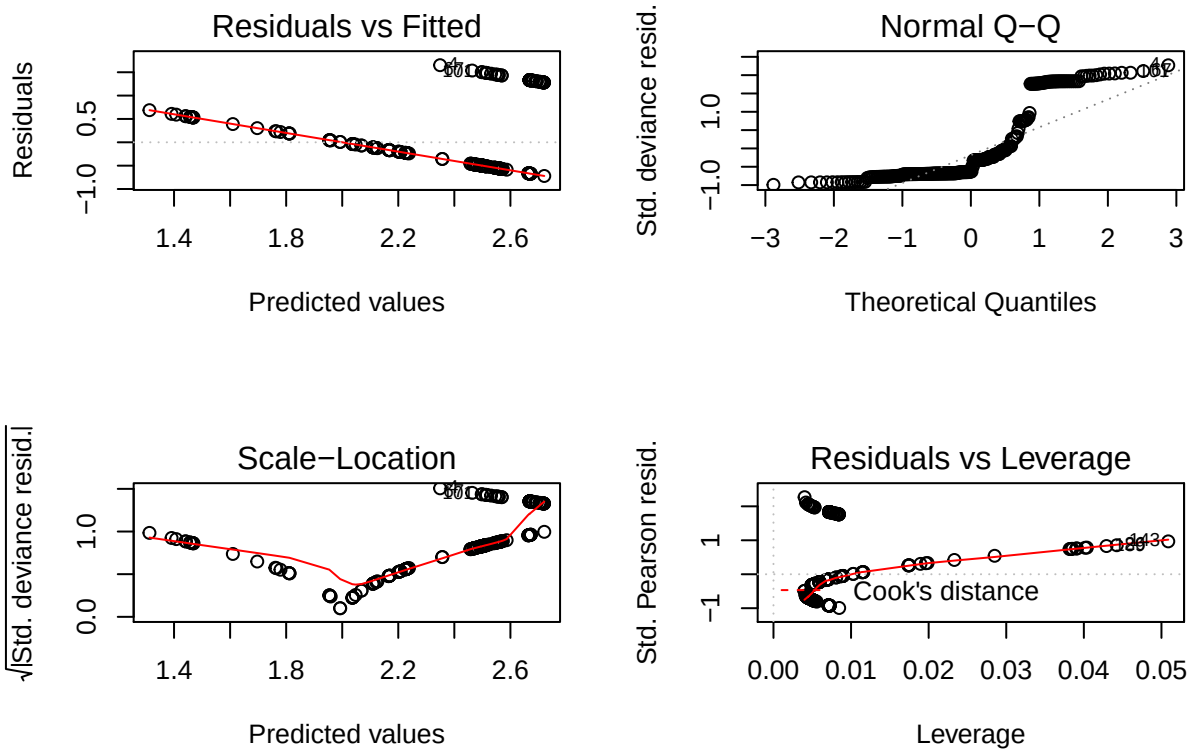
Regressione Lineare

Il primo modello che utilizziamo è la regressione lineare tra LSTAT e le 3 Principal Components.

```
##
## Call:
## lm(formula = BBcompressed$Poverta ~ ., data = BBcompressed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.566  -3.625  -0.713   2.480  20.763
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.6530632  0.2504129  50.529   <2e-16 ***
## comp1        0.0137842  0.0009573  14.398   <2e-16 ***
## comp2       -0.0156125  0.0015703  -9.942   <2e-16 ***
## comp3        0.0058921  0.0031705   1.858   0.0637 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.633 on 502 degrees of freedom
## Multiple R-squared:  0.3815, Adjusted R-squared:  0.3778
## F-statistic: 103.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

Da questi dati emerge che la terza componente, che prima avevamo incluso per poter arrivare a una percentuale prossima al 100%, in effetti è totalmente inutile nel predire la nostra feature LSTAT, quindi andrà scartata nei prossimi studi.

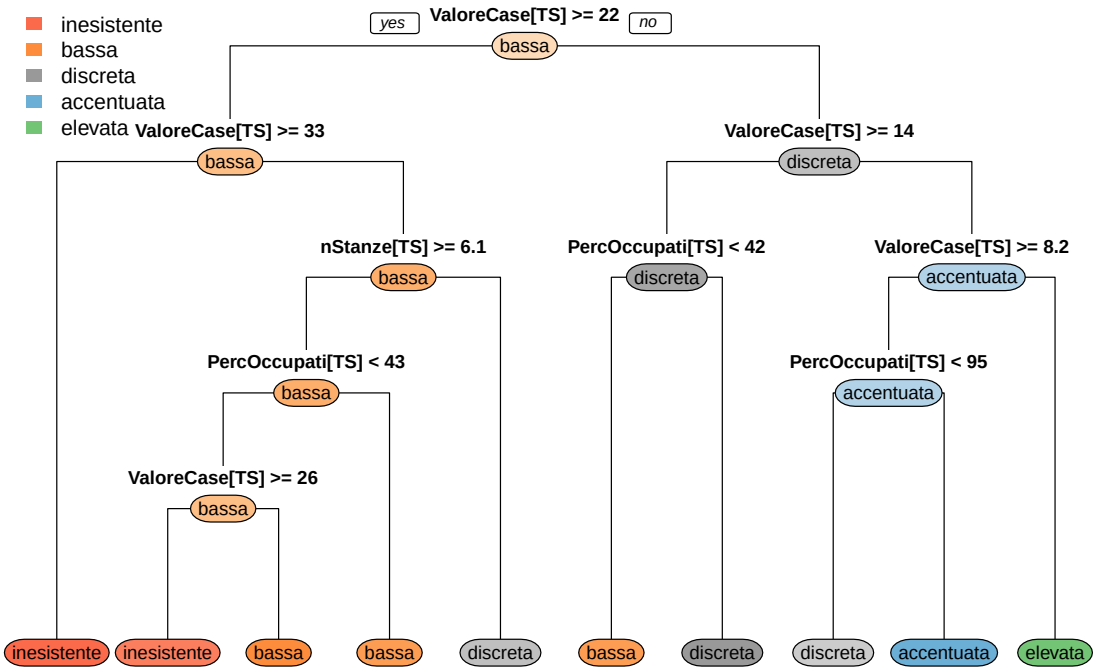
Regressione Logistica



La regressione logistica ci da delle informazioni riguardo due livelli di povertà, bassa e accentuata. Possiamo vedere Nel primo grafico come siano lineare e con pochi "outliers". Nel quarto grafico invece notiamo sempre una linearità tra i dati con qualche osservazione in alto a destra che si discosta.

Albero decisionale

Con questo terzo metodo vogliamo prevedere i 5 livelli di povertà sfruttando le feature: MEDV, RM e AGE. Costruiamo quindi un training set - TS e un control set - CS di 253 osservazioni ciascuno, infine applichiamo il DT.



Il plot prodotto sopra è la rappresentazione grafica dell'albero decisionale per prevedere la percentuale di povertà. Di seguito riportiamo la matrice di confusione sul Control Set.

```
##              BB.LivelloPoverta.CS.
## dataTree      inesistente bassa discreta accentuata elevata
##   inesistente           8    14       17          4       1
##   bassa                 7    30       21         12       5
##   discreta             12    52       33          9       4
##   accentuata            1     6        8          0       2
##   elevata               1     4        1          1       0
```

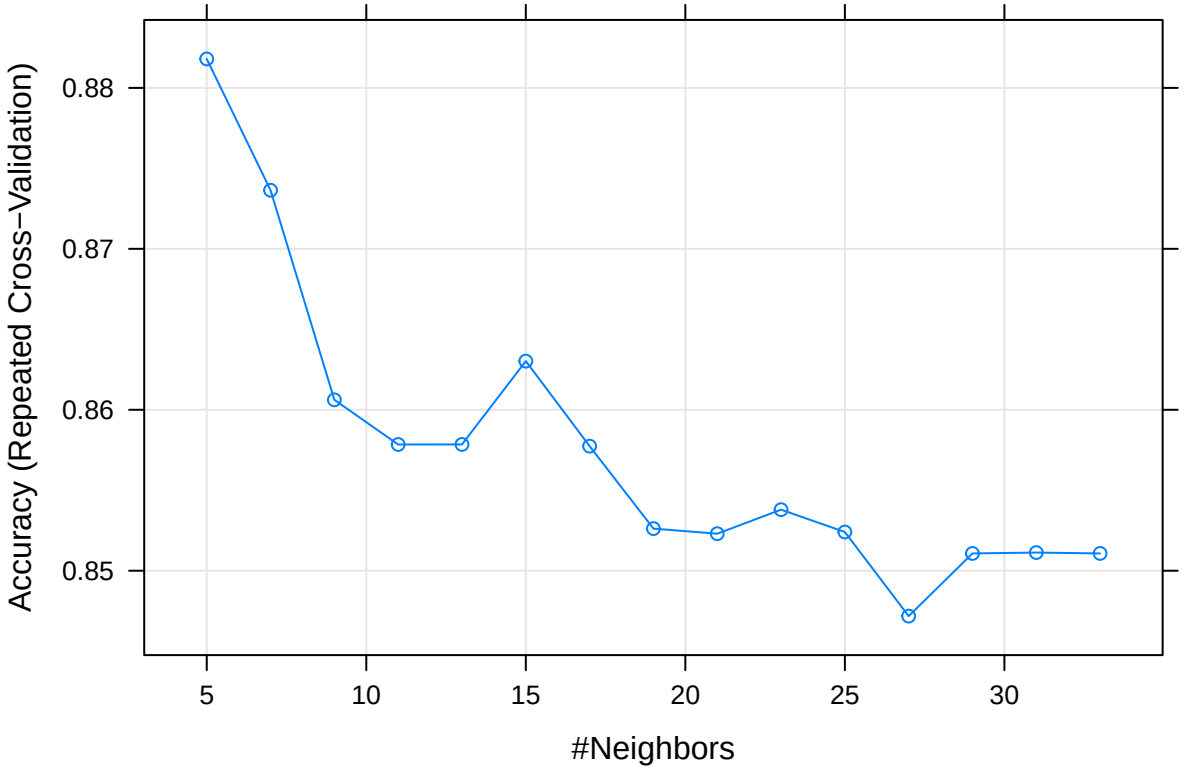
La predizione porta ad una discreta matrice di confusione. Troviamo punti di picco nelle colonne di inesistente, bassa e discreta povertà.

Algoritmo K-NN

Un altro metodo che abbiamo utilizzato per classificare la percentuale di povertà è quello dell'algoritmo K-NN. Le prove sono state fate per k=1,5,7.

La matrice di confusione dell'algoritmo K-NN è di certo migliore del modello dell'albero decisionale. Il risultato riportato sopra si riferisce a k=7. Abbiamo notato che aumentando i k, la matrice tende ad avere i risultati maggiori lungo la diagonale.

In basso possiamo apprezzare la rappresentazione grafica del risultato dell'algoritmo.



L'accuratezza del nostro algoritmo è abbastanza elevata e si avvicina verso lo 0.90% con k=15. Tende ad abbassarsi anche se rimane pressochè costante prendendo k =[17,35].

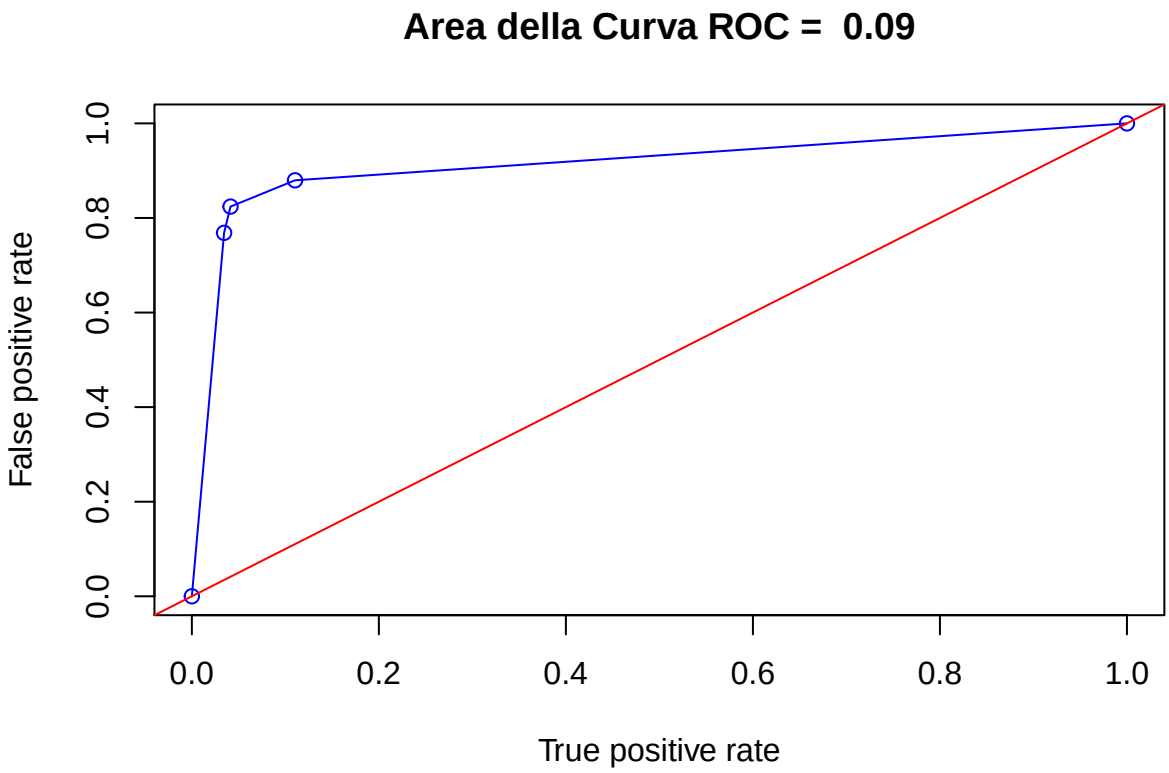
Curva ROC - Qualità del classificatore

In questo capitolo verifichiamo la bontà dei 2 modelli utilizzati precedentemente. Per facilitare i calcoli si è scelta una soglia di povertà pari al 13%, e abbiamo aggiunto una colonna che definisce questa separazione tra i livelli:

- Livello di Povertà $>$ 13% : povertà non trascurabile
- Livello di Povertà $<$ 13% : povertà trascurabile

Bontà del Decision Tree

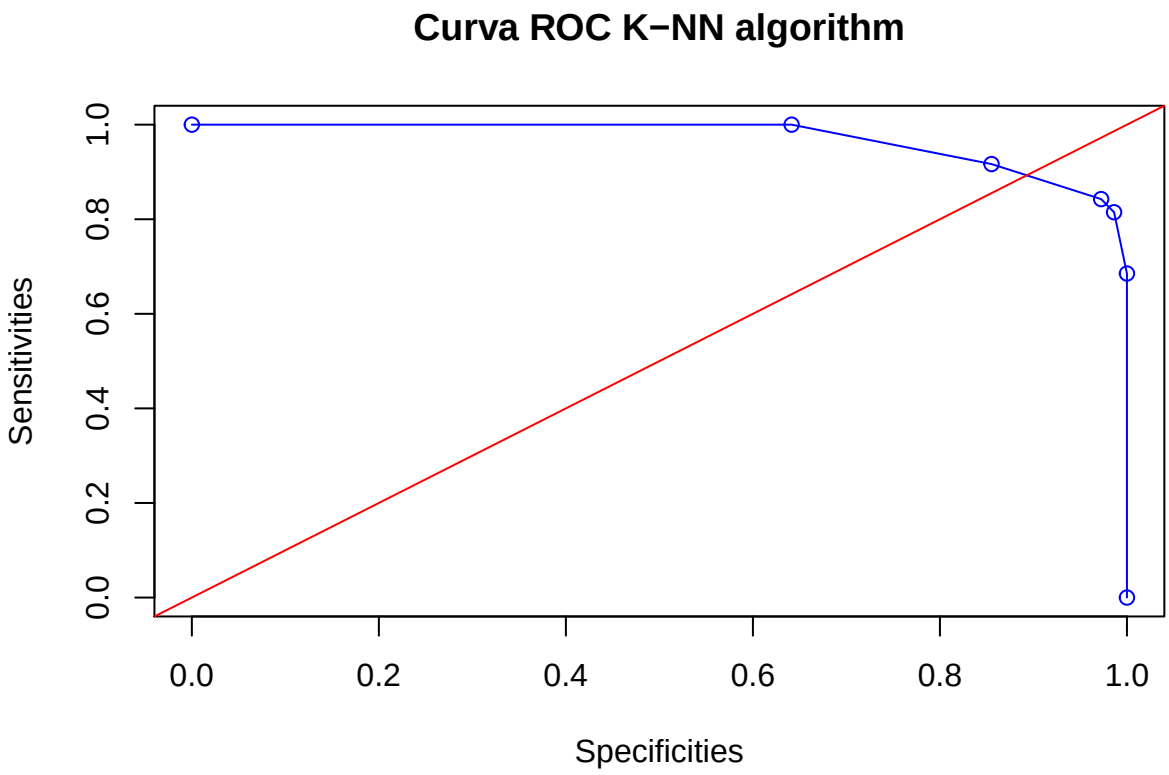
Studio della bontà dell'albero decisionale.



Da questo grafico possiamo notare che la nostra scelta di classificare il livello di povertà scegliendo le 3 features ValoreCase, PercOccupati e nStanze, produce un'ottima curva ROC, infatti possiamo notare che l'area che si crea sopra la diagonale è sufficientemente grande.

Bontà del K-NN

Infine studiamo la bontà dell'algoritmo K-NN.



In questo plot appena prodotto possiamo notare che la Curva ROC con le due grandezze inversamente proporzionali, Sensitivities e Specificities, danno, come punto di massimo, un risultato di 0.87. Questo valore da un peso ben specifico al nostro classificatore, infatti esso è ben bilanciato sia sulla specificità dei dati che osserva, sia sull'accuratezza degli stessi.

Conclusioni

In conclusione, il nostro classificatore che si pone di etichettare le osservazioni in base alla percentuale di povertà, si comporta molto bene sia utilizzando le Principal Components sia attraverso l'algoritmo KNN. Un risultato inaspettato è stato registrato con l'utilizzo dell'albero decisionale, che prima, osservando la matrice di confusione, ha prodotto dei valori "outliers" e discordanti, poi con la prova della curva ROC è stato valutato positivo e attendibile, anche se il valore della bontà rimane minore rispetto a quello registrato dall'algoritmo K-NN.

Il dataset Boston, composto da 13 features, può essere spiegato utilizzando solo 3 o addirittura 2 di queste:

- Valore medio delle case
- Percentuale delle case costruite prima del 1940 e occupate dagli stessi proprietari

I livelli di povertà maggiori che abbiamo registrato sono pressochè bassi. Il maggiore rilevato è quello di "bassa povertà", valore compreso tra il 6% e il 12%.