**Heart attack regression**

## Esame di machine learning 2021/2022

## Matteo Foschi

## Matricola: 535207

Sommario

[Sommario 2](#_Toc110431742)

[1. Dataset 3](#_Toc110431743)

[1.1 Features 3](#_Toc110431744)

[1.1 Training/Test set 3](#_Toc110431745)

[1.2 Standard scaler 3](#_Toc110431746)

[1.3 Riduzione dimensionalità PCA 4](#_Toc110431747)

[2 Modelli 4](#_Toc110431748)

[2.1 I modelli utilizzati sono 4](#_Toc110431749)

[2.1.1 Modello di linear regression 4](#_Toc110431750)

[2.1.2 Grafico performance 5](#_Toc110431751)

[2.1.3 Dati 5](#_Toc110431752)

[2.2.1 Modello logistic regession 6](#_Toc110431753)

[2.2.2 Grafico performance 7](#_Toc110431754)

[2.2.3 Dati 7](#_Toc110431755)

[2.3.1 La costo ridge 9](#_Toc110431756)

[2.3.2 Grafico performance 10](#_Toc110431757)

[2.3.3 Dati 10](#_Toc110431758)

[2.4.1 Lasso 12](#_Toc110431759)

[2.4.2 Grafico performance 13](#_Toc110431760)

[2.4.3 Dati 13](#_Toc110431761)

[2.5.1 La SVC 15](#_Toc110431762)

[2.5.2 Grafico performance 16](#_Toc110431763)

[2.5.3 Dati 16](#_Toc110431764)

[3 Obiettivo 19](#_Toc110431765)

[4 Considerazioni finali 19](#_Toc110431766)

[5 Link codice 19](#_Toc110431767)

# Dataset

## Features

Innanzitutto i dati che sono stati utilizzati per condurre sia l’addestramento che il test dei modelli, sono stati reperiti da Kaggle.

Il dataset è composto da 304 esempi, contenenti 13 features (features: [x1,x2,..]), più il valore della label di ogni esempio (target: y).

Le features presenti sono:

* età
* sesso
* tipo di dolore toracico (4 valori)
* pressione arteriosa a riposo
* colesterolo sierico in mg/dl
* glicemia a digiuno > 120 mg/dl
* risultati elettrocardiografici a riposo (valori 0,1 ,2)
* frequenza cardiaca massima raggiunta
* angina indotta dall'esercizio
* oldpeak = depressione del tratto ST indotta dall'esercizio rispetto al riposo
* la pendenza del segmento ST da esercizio di picco
* numero di vasi principali (0-3) colorati dalla fluoroscopia
* thal: 0 = normale; 1 = difetto fisso; 2 = difetto reversibile
* target: 0= meno possibilità di infarto 1= più possibilità di infarto

## Training/Test set

I dati sono stati elaborati prima di essere dati in pasto ai nostri modelli, tra cui una prima manipolazione è stata la suddivisione in due sotto insiemi di esempi, con esattezza:

* un training set contenente l’80% degli esempi
* un test set con il restante 20% dei dati

## Standard scaler

Poi è stata eseguita una standardizzazione delle funzionalità, rimuovendo la media e ridimensionando in base alla varianza unitaria.

È stata calcolata la media e la deviazione standard poi adattati i valori appena calcolati, sia ai miei dati di training che di test, restituendomi i dati trasformati.

## Riduzione dimensionalità PCA

L’ultima manipolazione ai dati che è stata eseguita in maniera dinamica, è stata la riduzione di dimensionalità mediante la PCA.

La riduzione di dimensionalità viene utilizzata, fissando una percentuale di informazione che si vuole mantenere, nel nostro caso siamo andati in un range tra 90% e 75% questo per vedere il comportamento anche dove la conservazione di informazione nei dati era più bassa.

# Modelli

## I modelli utilizzati sono:

* Modello di linear regression
* Modello di logistic regression
* Modello linear regression con costo ridge
* Modello linear regression con lasso
* Modello linear SVC con l2 penalty

### Modello di linear regression

Il modello lineare è l’unico modello che non ha subito tuning dei parametri.

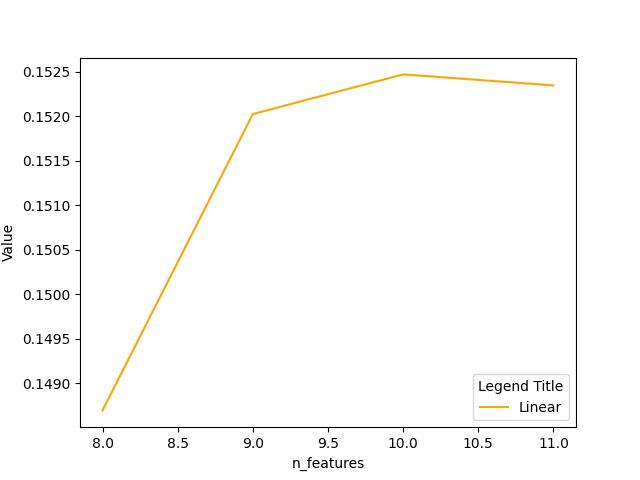
È stato eseguito solo mediante l’applicazione della riduzione della dimensionalità, si può vedere come il passaggio da un 75% ad un 90% di conservazione delle informazioni ha fatto sì che il valore della nostra regressione andasse via via aumentando. Con la linear regression non abbiamo ottenuto valori abbastanza elevati in quanto la funzione fosse costituita da solo due componenti per la generazione di un modello lineare (retta).

Come si vede dal grafico sotto man mano che le informazioni contenute nel training set diminuivano, andavano a diminuire anche le prestazioni sul test set. Il grafico rappresenta le prestazioni al test set e non al training set, anche se nei dati sotto al grafico possiamo osservare che anche l’accuracy ottenuta sul training set non è molto elevate ma sicuramente maggiore di quella generata dal test set.

Come si evince dal grafico al di sotto dell’80% delle informazioni il modello è peggiorato in maniera molto drastica.

Si può vedere chiaramente che il modello è in underfitting.

### Grafico performance:



### Dati:

---------Numero dimensionalita' usate 11 con una conservazione del 90.0% dell'informazione -------------

MODELLO LINEARE

Accuracy score test linear reg: 0.1523475390612956

Accuracy score training linear reg: 0.5196613851434475

---------Numero dimensionalita' usate 10 con una conservazione del 87.0% dell'informazione -------------

MODELLO LINEARE

Accuracy score test linear reg: 0.15247164576072014

Accuracy score training linear reg: 0.5196441102700324

---------Numero dimensionalita' usate 9 con una conservazione del 84.0% dell'informazione -------------

MODELLO LINEARE

Accuracy score test linear reg: 0.1520259749620212

Accuracy score training linear reg: 0.5158362278745904

---------Numero dimensionalita' usate 9 con una conservazione del 81.0% dell'informazione -------------

MODELLO LINEARE

Accuracy score test linear reg: 0.1520259749620212

Accuracy score training linear reg: 0.5158362278745904

---------Numero dimensionalita' usate 8 con una conservazione del 78 % dell'informazione -------------

MODELLO LINEARE

Accuracy score test linear reg: 0.14869548091127638

Accuracy score training linear reg: 0.5112743894345062

## Modello logistic regession

La logistic regression è stato attuato un tuning del parametro C in 4 step diversi, da 0.1 a 100 per vedere come variasse il comportamento del modello.

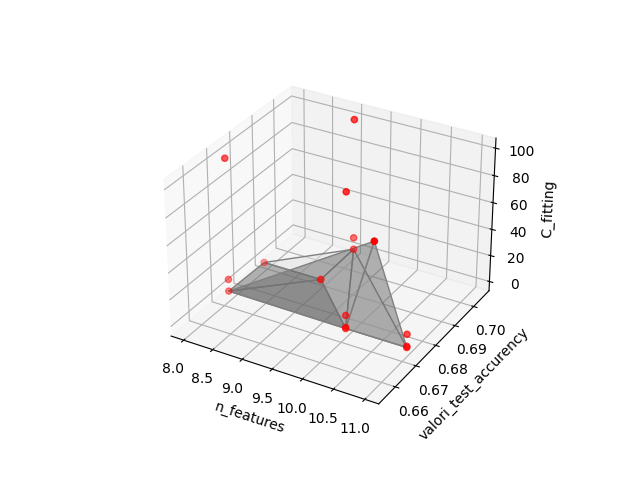
La visualizzazione questa volta è stata eseguita in 3D in quanto le prestazioni del nostro grafico variano: per ogni valore della PCA è stato eseguito il tuning in 4 valori differenti (0.1,1,10,100).

Dal grafico, quello che possiamo vedere è che ha dato dai valori molto stabili, in un range più limitato di accuracy, nonostante la riduzione mediante la PCA. Possiamo vedere come il valore sia anche aumentato, per poi calare nuovamente intorno all’80% dell’informazione.

Mentre nei primi casi aumentando il valore del parametro C abbiamo visto diminuire l’accuracy, viceversa alla riduzione dell’informazione nel training set, l’aumentare del valore di C ha migliorato le performance del modello.

Il modello ha raggiunto un buon addestramento, ma c’era ancora dal margine di miglioramento.

### Grafico performance:



### Dati:

---------Numero dimensionalita' usate 11 con una conservazione del 90.0% dell'informazione -------------

LOGISTIC REGRESSION

C: 0.1

Accuracy score test logistic reg: 0.6721311475409836

Accuracy score training logistic reg: 0.8388429752066116

C: 1.0

Accuracy score test logistic reg: 0.6721311475409836

Accuracy score training logistic reg: 0.8388429752066116

C: 10.0

Accuracy score test logistic reg: 0.6721311475409836

Accuracy score training logistic reg: 0.8347107438016529

C: 100.0

Accuracy score test logistic reg: 0.6557377049180327

Accuracy score training logistic reg: 0.8347107438016529

---------Numero dimensionalita' usate 10 con una conservazione del 87.0% dell'informazione -------------

LOGISTIC REGRESSION

C: 0.1

Accuracy score test logistic reg: 0.6721311475409836

Accuracy score training logistic reg: 0.8429752066115702

C: 1.0

Accuracy score test logistic reg: 0.6721311475409836

Accuracy score training logistic reg: 0.8305785123966942

C: 10.0

Accuracy score test logistic reg: 0.6721311475409836

Accuracy score training logistic reg: 0.8305785123966942

C: 100.0

Accuracy score test logistic reg: 0.6721311475409836

Accuracy score training logistic reg: 0.8305785123966942

---------Numero dimensionalita' usate 9 con una conservazione del 84.0% dell'informazione -------------

LOGISTIC REGRESSION

C: 0.1

Accuracy score test logistic reg: 0.6885245901639344

Accuracy score training logistic reg: 0.8512396694214877

C: 1.0

Accuracy score test logistic reg: 0.7049180327868853

Accuracy score training logistic reg: 0.8512396694214877

C: 10.0

Accuracy score test logistic reg: 0.7049180327868853

Accuracy score training logistic reg: 0.8471074380165289

C: 100.0

Accuracy score test logistic reg: 0.7049180327868853

Accuracy score training logistic reg: 0.8471074380165289

---------Numero dimensionalita' usate 9 con una conservazione del 81.0% dell'informazione -------------

LOGISTIC REGRESSION

C: 0.1

Accuracy score test logistic reg: 0.6885245901639344

Accuracy score training logistic reg: 0.8512396694214877

C: 1.0

Accuracy score test logistic reg: 0.7049180327868853

Accuracy score training logistic reg: 0.8512396694214877

C: 10.0

Accuracy score test logistic reg: 0.7049180327868853

Accuracy score training logistic reg: 0.8471074380165289

C: 100.0

Accuracy score test logistic reg: 0.7049180327868853

Accuracy score training logistic reg: 0.8471074380165289

---------Numero dimensionalita' usate 8 con una conservazione del 78% dell'informazione -------------

LOGISTIC REGRESSION

C: 0.1

Accuracy score test logistic reg: 0.6885245901639344

Accuracy score training logistic reg: 0.8471074380165289

C: 1.0

Accuracy score test logistic reg: 0.6721311475409836

Accuracy score training logistic reg: 0.8429752066115702

C: 10.0

Accuracy score test logistic reg: 0.6721311475409836

Accuracy score training logistic reg: 0.8471074380165289

C: 100.0

Accuracy score test logistic reg: 0.6721311475409836

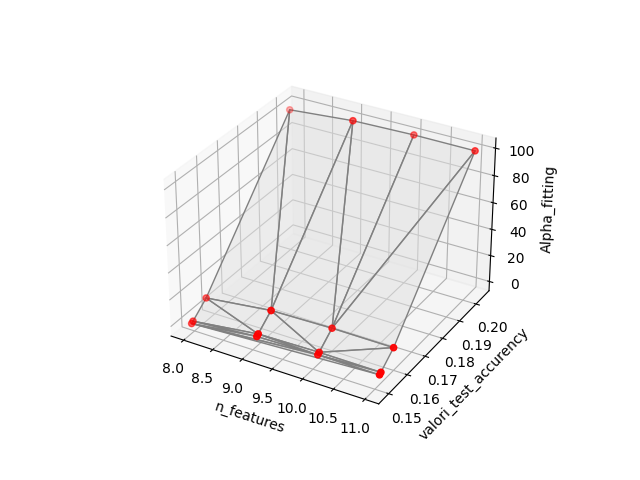
Accuracy score training logistic reg: 0.8471074380165289

## La costo ridge

La costo ridge come si può vedere dal grafico non ha raggiunto un livello di training elevati, tant’è che è chiaramente in underfitting il modello, ma il suo comportamento tende a migliorare con valori molto elevati di alpha, e tant’è che anche con la riduzione di informazioni nel training set i valori di accuracy sono diminuiti ma non in maniera così drastica. Il grafico mette in relazioni il comportamento del modello non solo al variare della percentuale dell’informazione nel training e test set, ma anche al variare di alpha nei 4 valori (0.1,1,10,100). Entrambe le variazioni hanno influito in maniera significativa sul comportamento del modello, ma provocato maggior variazione il parametro di tuning rispetto al quantitativo informativo dei dati.

Nonostante ciò, come vedremo più avanti, il suo comportamento è stato di gran lunga migliore rispetto alla lasso.

### Grafico performance:



### Dati:

---------Numero dimensionalita' usate 11 con una conservazione del 90.0% dell'informazione -------------LINEAR REGRESSION COSTO RIDGE

ALPHA 0.1

Accuracy score test ridge reg: 0.1524260791809009

Accuracy score training ridge reg: 0.5196613656958629

ALPHA 1.0

Accuracy score test ridge reg: 0.1531295997992952

Accuracy score training ridge reg: 0.5196594502418266

ALPHA 10.0

Accuracy score test ridge reg: 0.15984661203590445

Accuracy score training ridge reg: 0.5194772197253336

ALPHA 100.0

Accuracy score test ridge reg: 0.20398494135789502

Accuracy score training ridge reg: 0.5072869135186839

---------Numero dimensionalita' usate 10 con una conservazione del 87.0% dell'informazione -------------LINEAR REGRESSION COSTO RIDGE

ALPHA 0.1

Accuracy score test ridge reg: 0.15255008313076546

Accuracy score training ridge reg: 0.5196440908329616

ALPHA 1.0

Accuracy score test ridge reg: 0.15325268657763647

Accuracy score training ridge reg: 0.5196421764051857

ALPHA 10.0

Accuracy score test ridge reg: 0.15996121991830203

Accuracy score training ridge reg: 0.5194600354546672

ALPHA 100.0

Accuracy score test ridge reg: 0.2040527622703966

Accuracy score training ridge reg: 0.507272959372736

---------Numero dimensionalita' usate 9 con una conservazione del 84.0% dell'informazione -------------LINEAR REGRESSION COSTO RIDGE

ALPHA 0.1

Accuracy score test ridge reg: 0.1521032904361631

Accuracy score training ridge reg: 0.5158362100000394

ALPHA 1.0

Accuracy score test ridge reg: 0.15279594133927032

Accuracy score training ridge reg: 0.5158344484755373

ALPHA 10.0

Accuracy score test ridge reg: 0.15941794241029228

Accuracy score training ridge reg: 0.5156659702025532

ALPHA 100.0

Accuracy score test ridge reg: 0.20325786615445107

Accuracy score training ridge reg: 0.5040460711350996

---------Numero dimensionalita' usate 9 con una conservazione del 81.0% dell'informazione -------------LINEAR REGRESSION COSTO RIDGE

ALPHA 0.1

Accuracy score test ridge reg: 0.1521032904361631

Accuracy score training ridge reg: 0.5158362100000394

ALPHA 1.0

Accuracy score test ridge reg: 0.15279594133927032

Accuracy score training ridge reg: 0.5158344484755373

ALPHA 10.0

Accuracy score test ridge reg: 0.15941794241029228

Accuracy score training ridge reg: 0.5156659702025532

ALPHA 100.0

Accuracy score test ridge reg: 0.20325786615445107

Accuracy score training ridge: 0.5040460711350996

---------Numero dimensionalita' usate 8 con una conservazione del 78% dell'informazione -------------LINEAR REGRESSION COSTO RIDGE

ALPHA 0.1

Accuracy score test ridge reg: 0.1487727900879552

Accuracy score training ridge reg: 0.5112743731603029

ALPHA 1.0

Accuracy score test ridge reg: 0.14946548902173906

Accuracy score training ridge reg: 0.5112727683775918

ALPHA 10.0

Accuracy score test ridge reg: 0.15609720749856104

Accuracy score training ridge reg: 0.5111184114970815

ALPHA 100.0

Accuracy score test ridge reg: 0.20038559397626343

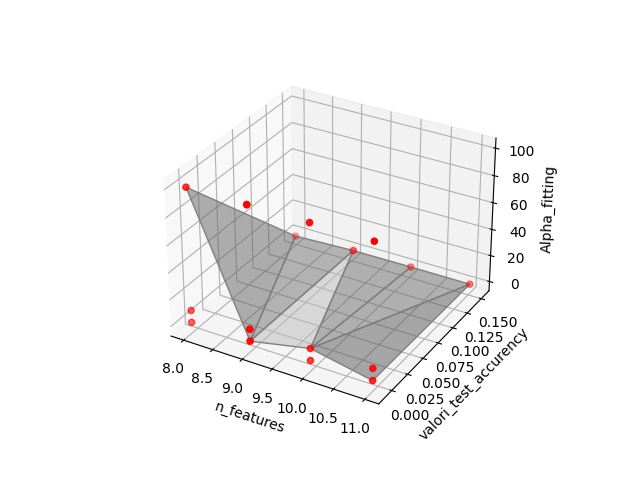
Accuracy score training ridge reg: 0.5001159038802286

## Lasso

Impiegando la lasso per il nostro dataset non ci ha dato valori per niente buoni, i valori “migliori” li abbiamo ottenuti impostando alpha ad un valore molto basso, ed all’aumentare di esso quel minimo di performance che avevamo raggiunto veniva compromesso. Anche in questo caso il grafico mostra la relazione sia al variare dell’informazione che di alpha. Alpha variava sempre nel range 0.1,1,10,100.

Come negli altri modelli il parametro alpha veniva fatto variare in ogni delta di variazione di informazione, il modello mostra anche con un’elevata percentuale di informazione un’underfitting.

### Grafico performance:



### Dati:

---------Numero dimensionalita' usate 11 con una conservazione del 90.0% dell'informazione -------------

ALPHA 0.1

Accuracy score test Lasso reg: 0.1523475390612945

Accuracy score training Lasso reg: 0.5196613851434474

ALPHA 1.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 10.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 100.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

---------Numero dimensionalita' usate 10 con una conservazione del 87.0% dell'informazione -------------LINEAR REGRESSION LASSO

ALPHA 0.1

Accuracy score test Lasso reg: 0.15247164576072048

Accuracy score training Lasso reg: 0.5196441102700322

ALPHA 1.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 10.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 100.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

---------Numero dimensionalita' usate 9 con una conservazione del 84.0% dell'informazione ------------- LINEAR REGRESSION LASSO

ALPHA 0.1

Accuracy score test Lasso reg: 0.15202597496202142

Accuracy score training Lasso reg: 0.5158362278745903

ALPHA 1.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 10.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 100.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

---------Numero dimensionalita' usate 9 con una conservazione del 81.0% dell'informazione ------------- LINEAR REGRESSION LASSO

ALPHA 0.1

Accuracy score test Lasso reg: 0.15202597496202142

Accuracy score training Lasso reg: 0.5158362278745903

ALPHA 1.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 10.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 100.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

---------Numero dimensionalita' usate 8 con una conservazione del 78% dell'informazione ------------- LINEAR REGRESSION LASSO

ALPHA 0.1

Accuracy score test Lasso reg: 0.1486954809112765

Accuracy score training Lasso reg: 0.5112743894345062

ALPHA 1.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 10.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

ALPHA 100.0

Accuracy score test Lasso reg: -0.008291342694289883

Accuracy score training Lasso reg: 0.0

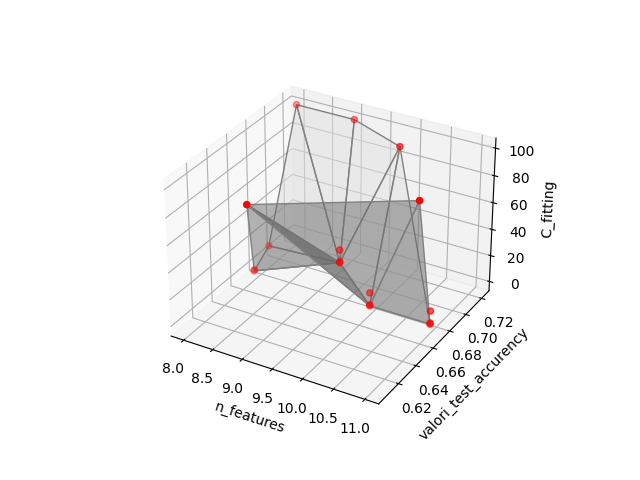
## La SVC

Il modello SVC è il modello che si è comportato migliore di tutti, con valori di accuracy mediamente buoni, anche al variare della percentuale di informazione nel training set. Tant’è che nei primi step con ha dato valori molto buoni che sono inizialmente migliorati all’aumentare di C, ma che successivamente sono diminuiti all’aumentare di C in maniera elevata.

Mentre con una percentuale di informazione più bassa, il parametro C non influenzava in maniera considerevole le perfomance, se non nei casi in qui C saliva a valori troppo alti. Nonostante siamo arrivati fino al 78% dell’informazione, su un numero ridotto a disposizione di esempi, con valori bassi di C, ha dimostrato ancora delle buone performance. Il modello ha raggiunto un discreto valore di addestramento ma ovviamente con un margine di miglioramento.

Il range di variazione di C è 01,1,10,100.

### Grafico performance:



### Dati:

---------Numero dimensionalita' usate 11 con una conservazione del 90.0% dell'informazione -------------SVC

C 0.1

Accuracy score test SVC reg L2: 0.6721311475409836

Accuracy score training SVC reg L2: 0.8388429752066116

C 1.0

Accuracy score test SVC reg L2: 0.6721311475409836

Accuracy score training SVC reg L2: 0.8429752066115702

C 10.0

Accuracy score test SVC reg L2: 0.6721311475409836

Accuracy score training SVC reg L2: 0.8429752066115702

C 100.0

Accuracy score test SVC reg L2: 0.6557377049180327

Accuracy score training SVC reg L2: 0.8223140495867769

---------Numero dimensionalita' usate 10 con una conservazione del 87.0% dell'informazione -------------

SVC

C 0.1

Accuracy score test SVC reg L2: 0.6721311475409836

Accuracy score training SVC reg L2: 0.8347107438016529

C 1.0

Accuracy score test SVC reg L2: 0.6721311475409836

Accuracy score training SVC reg L2: 0.8347107438016529

C 10.0

Accuracy score test SVC reg L2: 0.6721311475409836

Accuracy score training SVC reg L2: 0.8347107438016529

C 100.0

Accuracy score test SVC reg L2: 0.7049180327868853

Accuracy score training SVC reg L2: 0.8347107438016529

---------Numero dimensionalita' usate 9 con una conservazione del 84.0% dell'informazione -------------

SVC

C 0.1

Accuracy score test SVC reg L2: 0.7049180327868853

Accuracy score training SVC reg L2: 0.8553719008264463

C 1.0

Accuracy score test SVC reg L2: 0.7049180327868853

Accuracy score training SVC reg L2: 0.8553719008264463

C 10.0

Accuracy score test SVC reg L2: 0.7049180327868853

Accuracy score training SVC reg L2: 0.859504132231405

C 100.0

Accuracy score test SVC reg L2: 0.7213114754098361

Accuracy score training SVC reg L2: 0.7892561983471075

---------Numero dimensionalita' usate 9 con una conservazione del 81.0% dell'informazione ------------- SVC

C 0.1

Accuracy score test SVC reg L2: 0.7049180327868853

Accuracy score training SVC reg L2: 0.8553719008264463

C 1.0

Accuracy score test SVC reg L2: 0.7049180327868853

Accuracy score training SVC reg L2: 0.8553719008264463

C 10.0

Accuracy score test SVC reg L2: 0.7049180327868853

Accuracy score training SVC reg L2: 0.859504132231405

C 100.0

Accuracy score test SVC reg L2: 0.6065573770491803

Accuracy score training SVC reg L2: 0.7644628099173554

---------Numero dimensionalita' usate 8 con una conservazione del 78% dell'informazione -------------

SVC

C 0.1

Accuracy score test SVC reg L2: 0.6721311475409836

Accuracy score training SVC reg L2: 0.8429752066115702

C 1.0

Accuracy score test SVC reg L2: 0.6721311475409836

Accuracy score training SVC reg L2: 0.8429752066115702

C 10.0

Accuracy score test SVC reg L2: 0.6885245901639344

Accuracy score training SVC reg L2: 0.8471074380165289

C 100.0

Accuracy score test SVC reg L2: 0.7213114754098361

Accuracy score training SVC reg L2: 0.756198347107438

## Obiettivo

Durante lo sviluppo del progetto di machine learning, ho comparato diversi modelli di regressione, a seguito del tuning di alcuni parametri per migliorare la qualità del modello e vedere quale fosse il migliore. L’obiettivo della regressione era di identificare in base a delle caratteristiche (features) degli esempi, se il soggetto era a rischio di infarto o meno. Ogni modello veniva addestrato con lo stesso training set e valutato con lo stesso test set.

Inoltre è stato valutato anche il valore di addestramento mediante il confronto dei due parametri di accuracy del training e test set.

## Considerazioni finali

Abbiamo visto che asseconda del modello utilizzato la riduzione di dimensionalità andava ad elaborare e ridurre i dati, riducendone anche le informazioni contenute, ma che a sua volta rendeva più semplice l’addestramento del modello. Alcuni di essi si sono comportati bene anche con valori non molto alti di informazione, e sono anche migliorati sotto il tuning di alcuni parametri.

Altri modelli diversamente non hanno mostrato affatto un buon risultato, e si sono dimostrati notevolmente in underfitting.

I migliori sono stati SVC e logistic Regression mentre i peggiori lasso e costo ridge, a seconda della configurazione PCA e parametro di tuning.

Il test di accuracy è stato eseguito sia sul training che sul test set per poter analizzare se un modello era in underfitting o in overfitting.

## Link codice

https://github.com/MatteoIng/machineLearning.git