Progetto Big Data e Business Intelligence

Ollari Ischimji Dmitri

26 marzo 2022

Indice

Sommario

In questo documento verrano descritti i ragionamenti, le tecniche e gli strumenti utilizzati per cercare di ottenere il miglior risultato possibile nel task di classificazione multi classe sul dataset heart disease.

1 Look at the bigger pictures

Questo dataset proviene dal **UCI Machine Learning Repository**, viene utilizzato per task di classificazione supervisionate multiclasse poichè presenta 76 features, di cui una viene usata come target.

Lo scopo di questa task di classificazione è quello di poter distinguere le 5 classi del dataset:

1. Sano
2. Malato 1
3. Malato 2
4. Malato 3
5. Malato 4

Dove i vari malati si suddividono in classi (dalla 1 alla 4) in funzione della gravità.
Per valutare l'accuratezza del modello utilizzato, essendo un dataset sbilanciato, verranno utilizzate:

- Precision
- Recall
- F-1 Score

Considerando la tipoliga di task, sarebbe preferibile aver il minor numero di falsi positivi possibili.
Nel peggiore dei casi sarebbe preferibile sarebbe preferibile mantenere un recall più alto a discapito del precision score.

2 Get the data

Features Numero	Nome	Descrizione
1	ID	Identificativo del paziente
2	CCF	Social Securty Number
3	AGE	Età
4	SEX	Sesso del paziente (1 = M, 0 = F)
5	PAINLOC	Posizione del dolore toracico (1 = substernal, 0 = altro)
6	PAINEXER	dolore al petto provocato da sforzo = 1, altro = 0
7	RELREST	dolore passato dopo il riposo = 1, no = 0
8	PNCADEN	summa di 5,6 e 7
9	CP	Tipologia di dolore al petto: • tipica angina(malattia cardiaca) = 1 • angina atipica = 2 • non angina = 3 • asintomatico = 4
10	TRESTBPS	Pressione a riposo
11	HTN	storia di ipertensione 1 = si, 0 = no
12	CHOL	colesterolo
13	SMOKE	1 = fumatore, 0 = non fumatore
14	CIGS	sigarette al giorno
15	YEARS	anni di fumo
16	FBS	zuccheri nel sangue a digiuno maggiori di $120 mg/dl$, 1 = vero, 0 = falso
17	DM	Storia di diabete 1 = si, 0 = no
18	FAMHIST	presenza in famiglia di patologie alle coronarie
19	RESTECG	Elettrocardiogramma a riposo, 0 = normale, 1 = anormale
20	EKGMO	Mese della lettura per ECG
21	EKGDAY	Giorno della lettura per ECG
22	EKGYR	Anno della lettura per ECG
23	DIG	Usato il "digitalis" durante ECG? 1 = si, 0 = no
24	PROP	Beta Blocker usato durante ECG, 1 = si, 0 = no
25	NITR	nitrato usato nell'ECG, 1 = si, 0 = no
26	PRO	Calcio bloccato nelle letture dell'ECG, 1 = si, 0 = no
27	DIURETIC	usato un diuretico, 1 = si, 0 = no
28	PROTO	Protocollo usato per l'esercizio(classe da 1 a 12)
29	THALDUR	durata esercizio in minuti
30	THALTIME	Tempo dove è stata rilevata una depressione
31	MET	metriche raggiunte
32	THALACH	Battiti masimi
33	TAHLREST	Battiti cuore a riposo
34	TPEAKBPS	Picco di pressione durante l'esercizio (parte 1)
35	TPEAKBPD	Picco di pressione durante l'esercizio (parte 2)
36	DUMMY	dummy
37	TRESTBPD	pressione a riposo
38	EXANG	Angina indotta dall'esercizio? 1 = si, 0 = no
39	XHYPO	Calo di pressione indotto dall'esercizio? 1 = si, 0 = no
40	OLDPEAK	calo di pressione indotto a riposo

Features Numero	Nome	Descrizione
41	SLOPE	inclinazioen(?!) del picco sotto esercizio: • 1 = upsloping • 2 = flat • 3 = downsloping
42	RLDV5	massima a riposo
43	RLDV5E	massima in sforzo
44	CA	Numero di vasi maggiori colorati dalla fluoroscopia
45	RESTCKM	Irrilevante?!
46	EXERCKM	Irrilevante?!
47	RESTEF	non capisco
48	RESTWM	Movimento a riposo anormale(classi da 0 a 3)
49	EXEREF	espulsione frazionata radionucolide dovuta a sforzo
50	EXERWM	Non capisco
51	THAL	scansione al "taglium" durante l'esercizio: • 3 = normale • 6 = difetto fisso • 7 = difetto reversibile
52	THALSEV	non usata
53	THALPUL	non usata
54	EARLPUL	non usata
55	CMO	mese della cateterizzazione cardiaca?!
56	CDAY	giorno della cateterizzazione cardiaca?!
57	CYR	anno della cateterizzazione cardiaca?!
58	NUM	Rappresenta la classe che identifica la presenza e lo stato della malattia(usato come Y)
59	LMT	
60	LADPROX	
61	LADDIST	
62	DIAG	
63	CXMAIN	
64	RAMUS	
65	OM1	
66	OM2	
67	RCAPROX	
68	RCADIST	
69	LVX1	
70	LVX2	
71	LVX3	
72	LVX4	
73	LVF	
74	CATHEF	
75	JUNK	

2.1 Organizzare i dati in file csv

Ogni singono esempio è suddiviso in dieci righe. Per prima cosa si è dovuto convertire i 4 file in 4 versioni .csv molto più maneggevoli per l'analisi.

I dati sensibili sono stati eliminati dai gestori del repopository del dataset.

3 Explore data

I 4 dataset hanno:

- 292 esempi per il dataset di Cleveland
- 294 esempi per il dataset dell'Ungaria
- 200 esempi per il dataset di Long Beach
- 123 esempi per il dataset della Svizzera

3.1 Dataset unico

Analizzando individualmente i dati, si è osservato che alcune classi sono sotto rappresentate, rendendo difficoltoso l'apprendimento della stessa e la sua generalizzazione.

Un esempio relativo a questo problema lo si riscontra nell'analisi dei pazienti sani nel dataset "long-beach".

La soluzione che si è scelto seguire è quella di effettuare l'encoding di una nuova features relativa al dataset di origine(**location**):

- 0 = Cleveland
- 1 = hungarian
- 2 = long-beach
- 3 = switzerland

Dopo aver trasferito i dati nel formato csv, come primo passo sono state aggiunte le label a tutti i dataset e il dataset è stato divisi in TRAIN e TEST.

Si è preferito fare una prima pulizia per eliminare tutti quei dati che il testo del dataset definiva irrilevanti o dummy:

- id
- ccf
- dummy
- restckm
- exerckm
- thalsev
- thalpul
- earlobe
- lvx1
- lvx2
- lvx3
- lvx4
- lvf
- cathef
- junk
- name

3.2 PCA

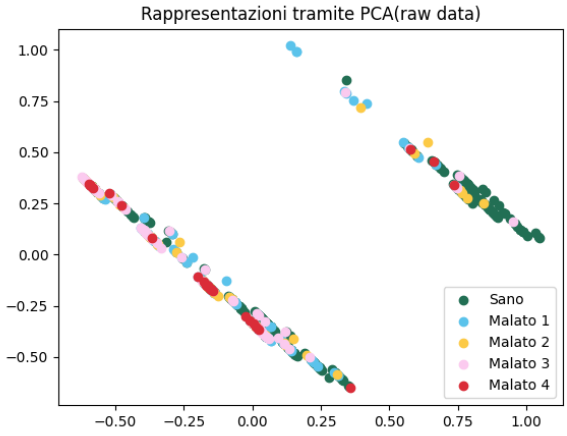


Figura 1: Data visualization con PCA

PCA è un metodo **non supervisionato** per effettuare riduzione dimensionale dei dati. Questa funzione è molto utile per permettere la rappresentazione di dati multidimensionali in uno spazio bidimensionale scegliendo i dati con maggior varianza.

3.3 LDA

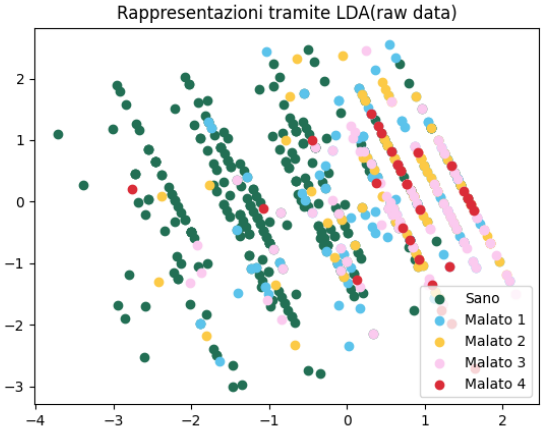


Figura 2: Data visualization con LDA

LDA è un metodo **supervisionato** per effettuare riduzione dimensionale dei dati.
Per effettuare la riduzione, crea un subset di features che rappresentano meglio la label target.

3.4 Considerazioni PCA e LDA

Entrambi i grafici fanno riflettere sul fatto che non siano corretti. LDA per esempio, presenta una strana separazione in 4 classi.

3.5 Correlazioni tra i dati

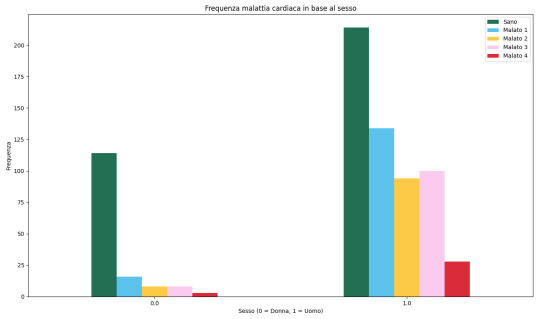


Figura 3: Correlazione salute-sesso

Da questo grafico si possono dedurre due considerazioni:

- Maggioranza di persone di sesso maschile prese in considerazione per l'analisi
- Maggior probabilità di malattia cardiaca per gli uomini

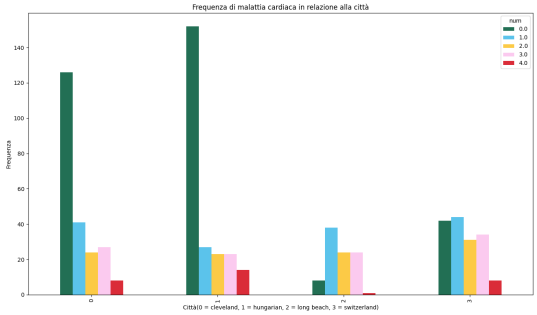


Figura 4: Correlazione salute-città

Da questo grafico si può capire il motivo per la realizzazione di un'unico dataset. Come rappresentato nel grafico, si hanno pochissime istanze in long beach e switzerland, la capacità di apprendere correttamente dal dataset ne risente nelle classi meno rappresentate.

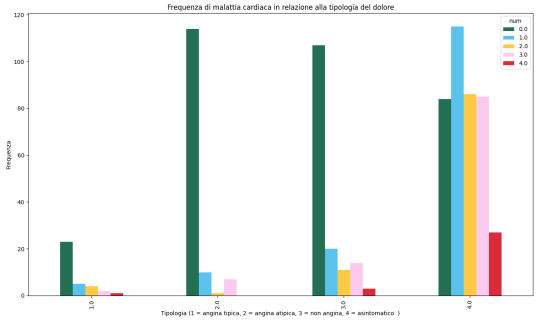


Figura 5: Correlazione salute-dolore

Mediante questo grafico possiamo intuire che la maggior parte dei pazienti malati non presenti sintomi fisici(dolore al petto) nella quotidianità.

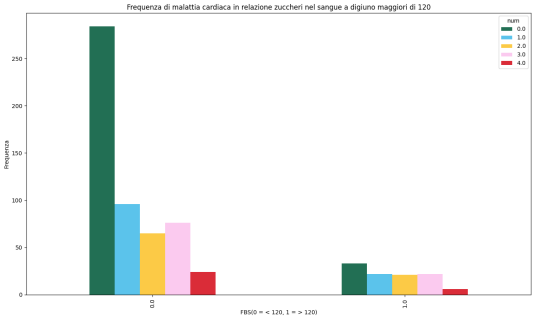


Figura 6: Correlazione salute-zuccheri

Solitamente con una densità di zuccheri a digiuno superiore a $120 mg/dL$ si viene considerati diabetici. Da questo grafico si può intuire che la percentuale di persone sottoposta all'analisi dei dati è prevalenetemente **non diabetica**. Qualora il paziente venga identificato come diabetico la probabilità di malattia cardiaca aumenta più del doppio rispetto ad un paziente non affetto dalla stessa.

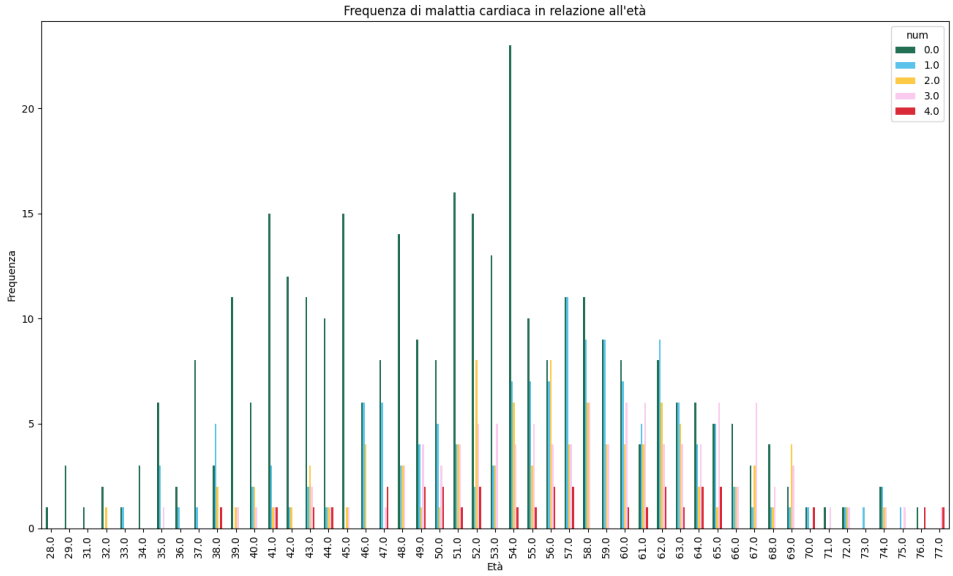


Figura 7: Correlazione salute-età

Con questo grafico si può visualizzare una distribuzione a campana dei pazienti in correlazione all'età.

4 Prepare data for ML algorithms

Questa è la parte più corposa del progetto poichè, in termini di codice, il dataset originale presenta 76 features di cui 1 classe target.

4.1 Non unique

In questa fase si sono cercate quelle features che non apportano informazioni, poichè costanti.

È stata rimossa la features **PNCADEN**.

4.2 Analisi dei valori nulli

Nel dataset i valori nulli sono salavati con il valore di -9. Quindi il primo passo è stato quello di sostituire i valori con **pd.nan** che restituisce il valore nullo.

Dopo aver ottenuto il dataset con i valori nulli, si è proseguito con il rimuovere tutte le colonne e tutte le righe contenenti una percentuale di valori nulli maggire del 15%.

4.3 Outliners

Per poter effettuare il processo di eliminazione degli outliners bisogna, per prima cosa, salvare i dati numerici in un array e successivamente analizzare la deviazione standard per includere solo il 99% dei dati.

Mediante la funzione **zscore** della libreria stats si sono eliminati gli esempi con una deviazione maggiore del 3.

4.4 Date

Studiando il problema si è pensato che le date non apportassero informazioni utili e si è preferito rimuoverle.

4.5 Variabile "proto"

La variabile proto è una features categorica con 12 casistiche e rappresenta la tipologia del protocollo utilizzato per il test da sforzo.

Dal'analisi di questa features si è notato che la maggior parte dei valori non sono coerenti con le 12 classi. Si è di conseguenza preferito rimuovere l'intera colonna della feature.

4.6 Fillna

Si è provveduto ad assegnare la tipolgia ai dati(float, category, ecc) e successivamente assegnare, ai dati ancora nulli il valore medio presente per la propria colonna nel dataset.

4.7 Features selection

Dato che alcune features presentano valori negativi non si è potuto ricorrere al chi2 per l'analisi delle migliori features.

Si è scelto di utilizzare due funzioni:

- ANOVA con la funione f classif
- mutaul info regression

Mutual info regression ha attribuito il risultato più basso alle seguenti features:

- xhypo
- htn
- diuretic
- trestbpd
- thalrest
- nitr
- tpeakbps
- prop
- pro

Anova ha attributo il risultato più basso alle seguenti features:

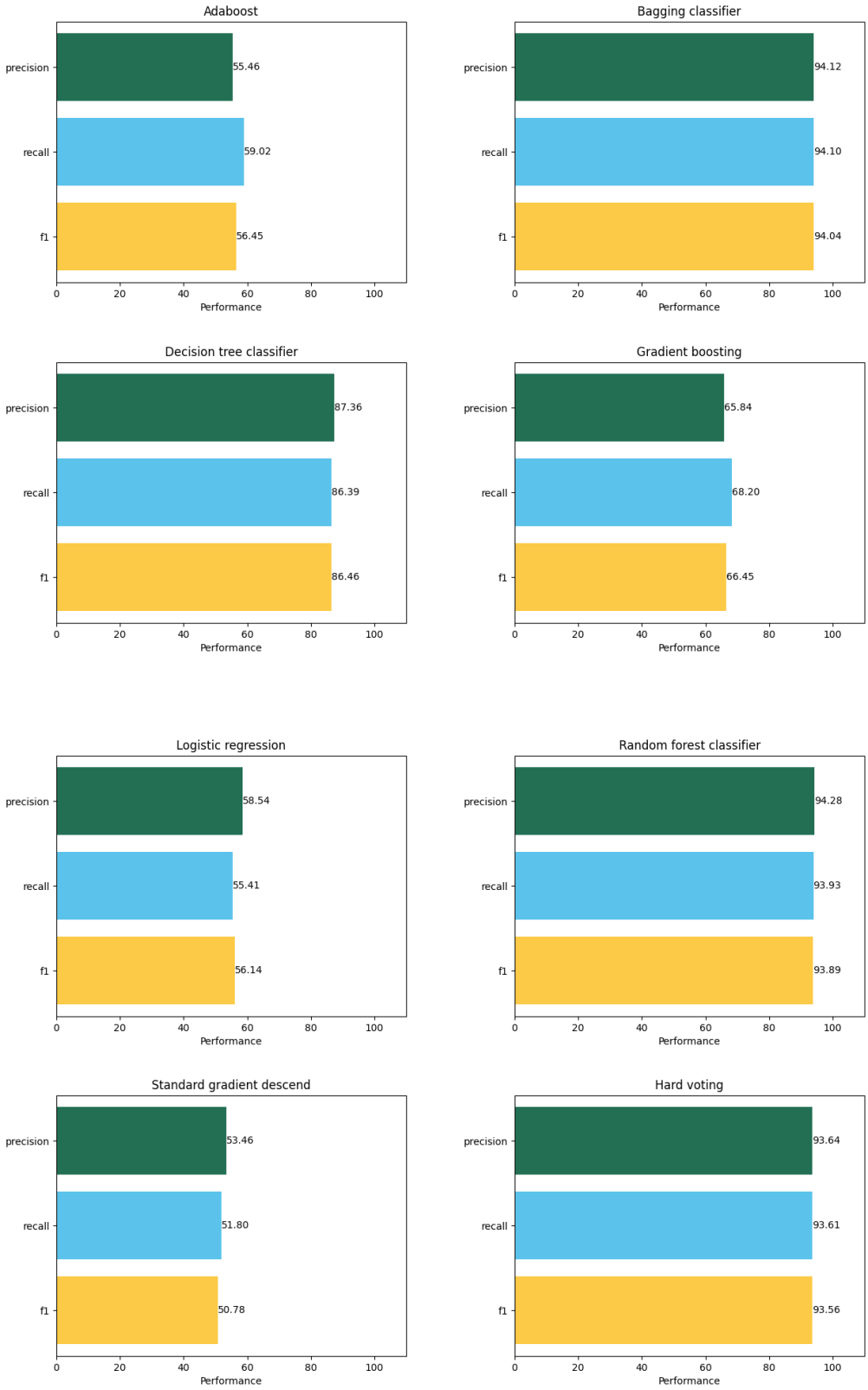
- tpeakbpd
- dig

Il risultato finale dopo il datacleaning è il seguente sub-set:

Nome	Descrizione
age	Età del soggetto
sex	Sesso del soggetto(0 = F, 1 = M)
cp	Tipologia di dolore al petto
trestbps	Pressione a riposo
chol	colesterolo
fbs	zuccheri nel sangue a digiuno maggiori di $120mg/dl$, 1 = vero, 0 = falso
restecg	Elettrocardiogramma a riposo, 0 = normale, 1 = anormale
thaldur	durata esercizio in minuti
met	metriche raggiunte
thalach	Battiti masimi
exang	Angina indotta dall'esercizio? 1 = si, 0 = no
oldpeak	calo di pressione indotto a riposo
location	variabile encodata per rappresentare il dataset di provenienza
target	Rappresenta la classe che identifica la presenza e lo stato della malattia(usato come Y)

5 Model comparison

Ad ogni modello è stato applicata una k-fold di 10 e sono state effettuate prove con il dataset ottenuto dalla features selection. Per lo stesso dataset sono state eseguite prove anche con un dataset bilanciato ottenuto tramite oversampling(smote).



Dopo un controllo con il train-set iniziale si è capito che smote non migliora i punteggi e di conseguenza si è preferito mantenere il dataset sbilanciato.

6 Evaluation

Ogni modello è stato "fittato" 10 volte mediante k-fold sia con dati sbilanciati che con dati bilanciati tramite over sampling.

Sono stati regolati i parametri di fine tuning e, una volta ottenuto il risultato migliore, si è provveduto a fare la prova finale con il test-set separato all'inizione del processo di machine learning.

Considerata la natura del test, cioè dover prevedere se un paziente sia malato o meno, si è preferito dare leggermente maggior importanza al punteggio di **recall**, essendo il recall il valore che rappresenta la misura dell'errore sui "false positive".

Pertanto i modelli migliori che risultano da questa analisi sono:

- Bagging classifier
  - precision = 94.12%
  - recall = 94.10%
  - f-1 = 94.04%
- Random forest classifier
  - precision = 94.28%
  - recall = 93.93%
  - f-1 = 93.89%
- Hard voting
  - precision = 93.64%
  - recall = 93.61%
  - f-1 = 93.56%

7 Fonti

Il dataset proviene da: UCI machine learning repository.

Creatori:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donatore:

- David W. Aha (aha '@' ics.uci.edu)