Politecnico di Torino
Laurea magistrale in Ingegneria Informatica

Tesina di Data Spaces

Docente:

Prof. Francesco Vaccarino

Studente:

Luca Vezzani

Anno accademico 2019/2020

Sommario

Introduzione all'argomento

Citando il **National Heart, Lung and Blood Institute:**

> "*Con il termine "malattia cardiaca" ci si riferisce ad un insieme di condizioni che influenzano la struttura e le funzioni del cuore. Ad esempio, la malattia coronarica è un tipo di malattia cardiaca che si sviluppa quando le arterie non riescono a trasportare abbastanza sangue ricco di ossigeno al cuore; questa è la principale causa di morte negli stati Uniti*."
>
> Fonte: https://www.nhlbi.nih.gov/health-topics/espanol/enfermedad-coronaria

In più, secondo la **World Health Organization**, le malattie coronariche sono la principale causa di morte nel mondo (Fonte: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

L'obiettivo di questa introduzione è dare abbastanza informazioni di base per comprendere il dataset usato per questa analisi, ovvero l' Heart Disease UCI, utilizzato per predire se un paziente (con determinati valori) presenta una malattia cardiaca o no.

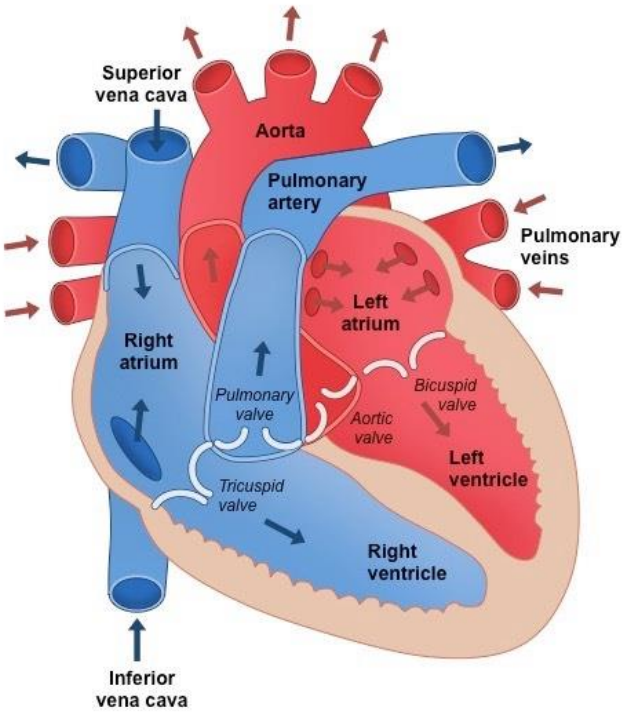
Contesto

Cuore e sangue

Il sangue è fondamentale per il corretto svolgimento di tutte le funzionalità del corpo; esso trasporta ossigeno e nutrienti a tutte le cellule e raccoglie tutti gli elementi di scarto da queste.

Il sangue riesce a raggiungere tutto il corpo perché pompato dal cuore: questo organo riceve il sangue povero di ossigeno, lo invia ai polmoni a far rifornimento e dopo lo pompa nuovamente verso il corpo.

L'immagine qui a destra rappresenta la struttura interna del cuore: le frecce blu rappresentano il sangue povero diretto ai polmoni, mentre le rosse quello di ritorno da essi e verso il resto del corpo.



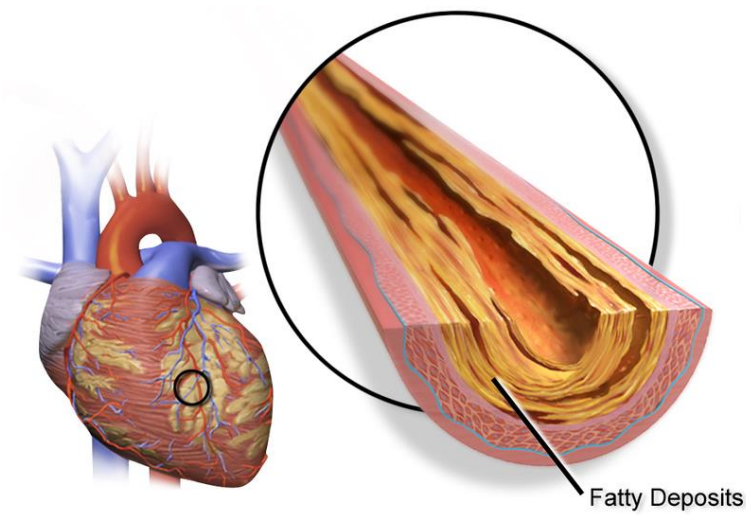
Una scorretta distribuzione del sangue nell'organismo impedisce il corretto svolgimento delle funzioni che, nel caso peggiore, può determinare la morte della cellula.

Malattie coronariche

Non bisogna dimenticare che, lo stesso cuore, per funzionare, necessita di nutrienti che vengono forniti da arterie conosciute come “coronarie”.

Quando si parla di malattia coronarica ci si riferisce spesso ad una difficoltà incontrata dal sangue nel fluire in queste arterie a causa di grasso accumulato al loro interno.

Nel caso peggiore, la conseguenza che si ha lasciando il cuore senza il giusto nutrimento porta al cosiddetto “attacco cardiaco” o, in altri termini, la morte di parte delle cellule del cuore.

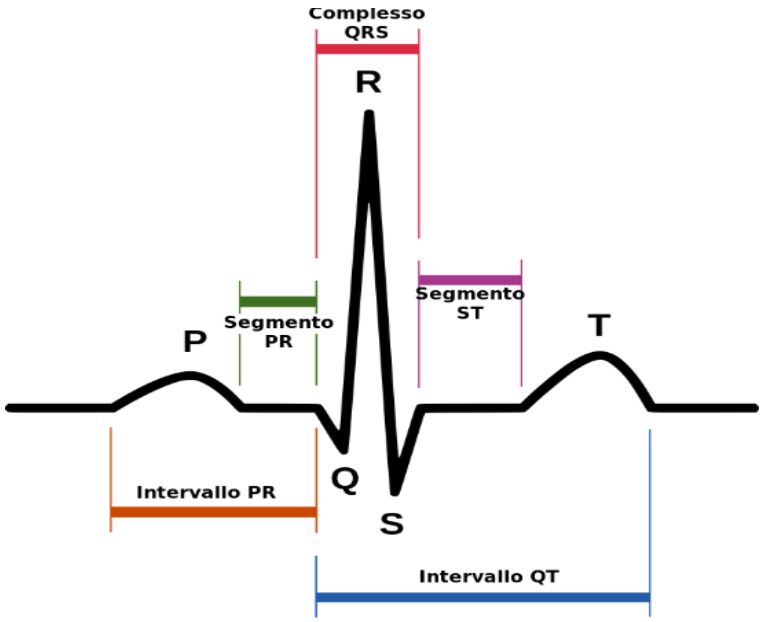


Glossario dei termini

- **Aterosclerosi:** accumulo di sostanze sulle pareti delle arterie che può alterare il normale flusso sanguigno. Inoltre, la rottura di questi ammassi può portare alla creazione di trombi che possono bloccare completamente il flusso.
- **Ischemia**: riduzione dell’apporto normale di sangue ad uno specifico tessuto, tra cui ossigeno e sostanze che possono portare al non funzionamento dell’apparato.
- **Angina**: dolore pettorale dovuto ad una riduzione del flusso sanguigno nelle arterie coronarie; sono presenti varie tipologie:
    - **Angina stabile**: causata da situazioni che richiedono ossigeno (esercizio o stress) e che scompare con il riposo;
    - **Angina instabile**: angina che può avvenire anche a riposo;
    - **Angina tipica**: angina che si mostra sotto forma di fastidio a petto;
    - **Angina atipica**: angina che si mostra sotto forma di nausea o mancanza di fiato.
- **Trombo:** massa di sangue allo stato solido che impedisce il flusso sanguigno in un canale.
- **Embolo:** trombo che si stacca dal canale di origine e viaggia in altre parti del corpo.
- **Infarto acuto del miocardio:** conosciuto anche come attacco di cuore, è la morte di parte delle cellule che compongo l’organo per mancanza di ossigeno.
- **Nuclear stress test:** test nel quale una sostanza radioattiva è iniettata al paziente per vedere il flusso del sangue a riposo e durante esercizio (accompagnata solitamente da un elettrocardiogramma).
- **Malattia asintomatica:** caso in cui il paziente risulta avere la malattia ma non ha sintomi (caso più difficile da gestire a livello di data modeling).

- **Ipertrofia ventricolare sinistra:** ispessimento delle pareti della camera principale del cuore che pompa il sangue al resto del corpo; questo può provocare un rilassamento del muscolo che non riesce, a lungo andare, a funzionare correttamente.

- **Elettrocardiogramma:** grafico dei segnali elettrici che causano il battito cardiaco; ogni parte del segnale ha un nome e quelle più interessanti per le malattie cardiache sono l'onda T e il segmento ST.



Dataset

Questo dataset è presente su Kaggle ed è stato fornito dalla **UCI Machine Learning Repository**. In esso si trovano circa 300 record di pazienti di Cleveland con le features descritte di seguito.

Errori nel dataset

Analizzando il dataset è possibile notare come alcuni valori siano errati (frutto di trascrizioni errate probabilmente):

- le righe # 93, 159, 164, 165 e 252 hanno un valore di ca = 4 che non esiste (nel dataset originale erano NaN);
- le righe #49 e 282 hanno thal = 0, anche questo errato (sempre NaN nel dataset originale).

Dato che sono un piccolo numero, ho deciso di eliminarle dal dataset.

Dataset features

Target

Forse il dato più importante, indica se il paziente è affetto da una malattia cardiaca o no:

- "0": si
- "1": no

Età (age)

Età del paziente in anni.

Sesso (sex)

Il sesso del paziente:

- "0": donna
- "1": uomo

Dolore al petto (cp)

Il tipo di dolore al petto che presenta il paziente:

- "0": asintomatica
- "1": angina atipica
- "2": dolore non collegato ad angina
- "3": angina tipica

Il modo in cui questo dato viene preso non è descritto.

Pressione del sangue a riposo (trestbps)

Pressione del sangue a riposo misurata in millimetri di mercurio (mm Hg) rilevata quando il paziente si è presentato in ospedale.

Colesterolo (chol)

Livello di colesterolo nel siero misurato in mg/dl.

Glicemia a digiuno > 120 mg/dl (fbs)

Valore che indica se la glicemia è superiore di un livello limite:

- "0": no
- "1": si

Elettrocardiogramma a riposo (restecg)

Risultato dell'ecg con il paziente a riposo

- "0" = probabile ipertrofia ventricolare sinistra
- "1" = normale
- "2" = anomalie nell'onda T o nel segmento ST

Massimo battito cardiaco raggiunto(thalach)

Massimo battito cardiaco raggiunto sotto stress test.

Presenza di angina durante esercizio (exang)

Denota la presenza/assenza di angina durante l'esercizio fisico

- "0": no
- "1": si

Depressione del segmento ST (oldpeak)

Valore che identifica se e di quanto il segmento ST diminuisce sotto sforzo rispetto al riposo.

Pendenza del segmento ST (slope)

Indica la pendenza del segmento sopracitato durante la maggior parte dell'esercizio fisico

- "0": discendente
- "1": piatto
- "2": ascendente

Numero di vasi principali colorati da fluoroscopia (ca)

Numero dei vasi principali colorati da una colorazione radioattiva. Il numero va da 0 a 3.

Difetto cardiaco (thal)

Risultato dell'osservazione del flusso sanguigno osservato tramite colorazione radioattiva.

- "1": difetto fisso (flusso non presente in alcune parti del cuore)
- "2": flusso normale
- "3": difetto reversibile (flusso presente ma non normale)

Esplorazione del dataset

Prima di procedere con l'applicazione delle varie tecniche, ho analizzato i dati per avere una buona visione d'insieme e per identificare gli attributi che più influiscono sul risultato.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Come descritto prima e come si può notare dall'immagine, il dataset è formato da 14 colonne che contengono dati numerici.

Al fine di rendere più comprensibili alcuni grafici, ho deciso di creare una copia del dataset e rinominare le variabili numeriche con la caratteristica associata nella descrizione, ottenendo

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	uomo	angina tipica	145	233	si	ipert. ventricolare	150	no	2.3	downslope	0	aggiustato	no
1	37	uomo	dolore no-angina	130	250	no	normale	187	no	3.5	downslope	0	normale	no
2	41	donna	angina atipica	130	204	no	ipert. ventricolare	172	no	1.4	upslope	0	normale	no
3	56	uomo	angina atipica	120	236	no	normale	178	no	0.8	upslope	0	normale	no
4	57	donna	asintomatico	120	354	no	normale	163	si	0.6	upslope	0	normale	no

sicuramente ad una prima occhiata risulta molto più comprensibile, non rendendo per forza necessaria la lettura della descrizione dei valori.

Attraverso il comando *describe* è possibile visionare alcune statistiche generali sul dataset:

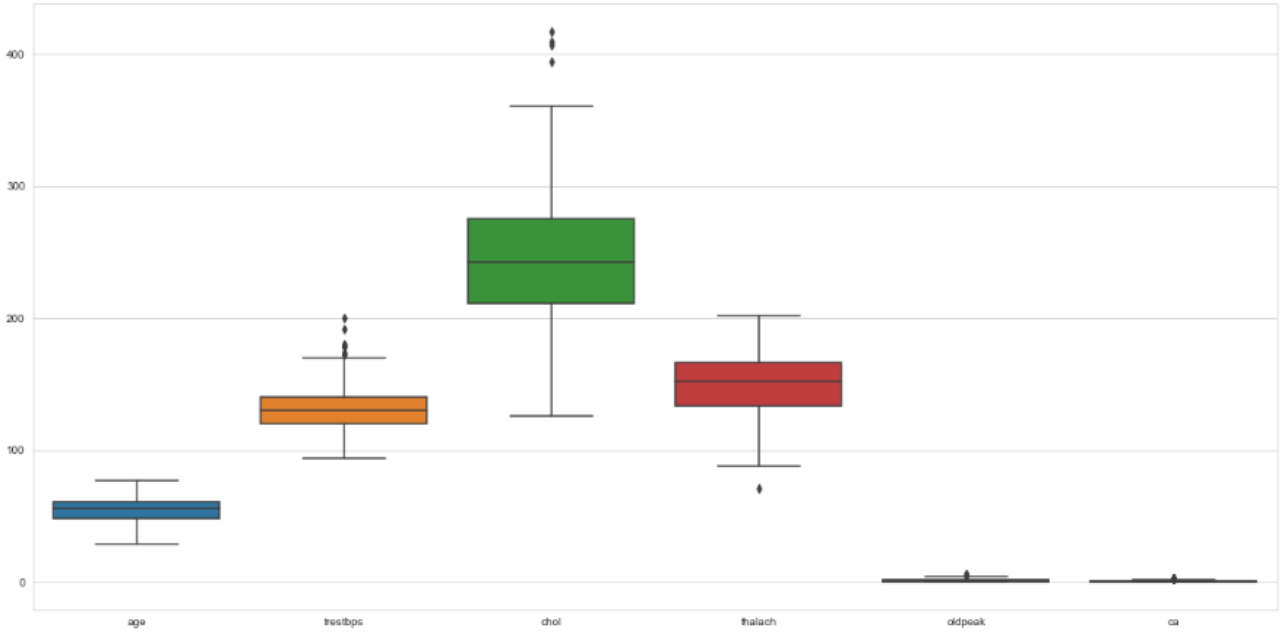
(non sono rappresentate le variabili categoriche anche se sotto forma numerica)

	age	trestbps	chol	thalach	oldpeak	ca
count	295.000000	295.000000	295.000000	295.000000	295.000000	295.000000
mean	54.481356	131.661017	246.081356	149.525424	1.057288	0.681356
std	9.045548	17.730225	48.663919	23.001742	1.168029	0.940487
min	29.000000	94.000000	126.000000	71.000000	0.000000	0.000000
25%	48.000000	120.000000	211.000000	133.000000	0.000000	0.000000
50%	56.000000	130.000000	242.000000	152.000000	0.800000	0.000000
75%	61.000000	140.000000	275.000000	166.000000	1.700000	1.000000
max	77.000000	200.000000	417.000000	202.000000	6.200000	3.000000

Qui possiamo analizzare:

- **Count:** numero totale di record presente nel dataset (nel nostro caso numero di pazienti)
- **Mean:** valore medio intorno al quale si espande l'attributo
- **Std:** deviazione standard per ogni gruppo di attributi (indice del grado di dispersione intorno alla media)
- **25%, 50%** e **75%:** percentili o quartili che indicano la percentuale sotto la quale si verificano una determinata percentuale di osservazioni (il 50% corrisponde con la mediana ma fornisce un'ulteriore informazione sulla distorsione della distribuzione)
- **Max/min:** valore maggiore/minore del gruppo

Rappresentando questi valori attraverso boxplot:



possiamo notare come la pressione del sangue a riposo (trestbps) e il colesterolo nel siero (chol) presentino degli outliers, ovvero valori che vanno al di fuori della deviazione standard dei valori di quella features, ma possiamo notare che sono pochi e non stravolgono quest'ultima.

Analisi features:

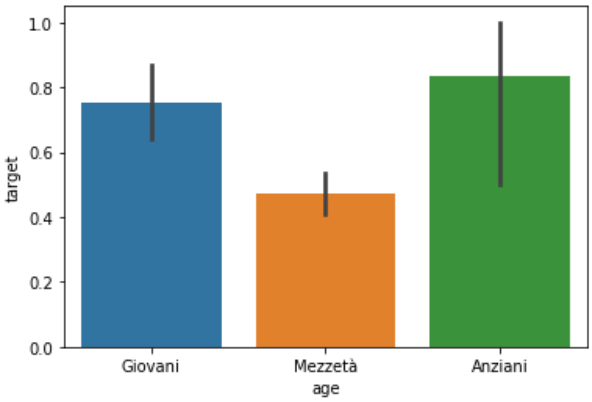
Per identificare i fattori di rischio per le malattie cardiache, ho esaminato ogni feature presente nel dataset confrontandolo con il "target" ovvero i pazienti affetti o no da malattia attraverso barplot, countplot e boxplot a seconda delle necessita.

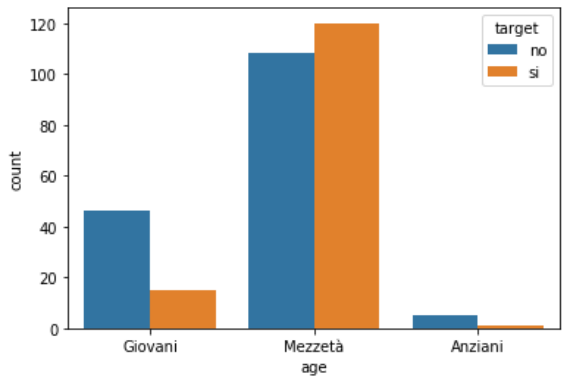
Età:

Suddivisione categorie:

- Giovani < 45
- 45 < Mezz'età < 70
- Anziani > 70

Il barplot mostra un risultato abbastanza strano: se a primo acchito, esaminando giovani e pazienti di mezz'età, può sembrare che l'età sia un fattore di rischio per le malattie cardiache,

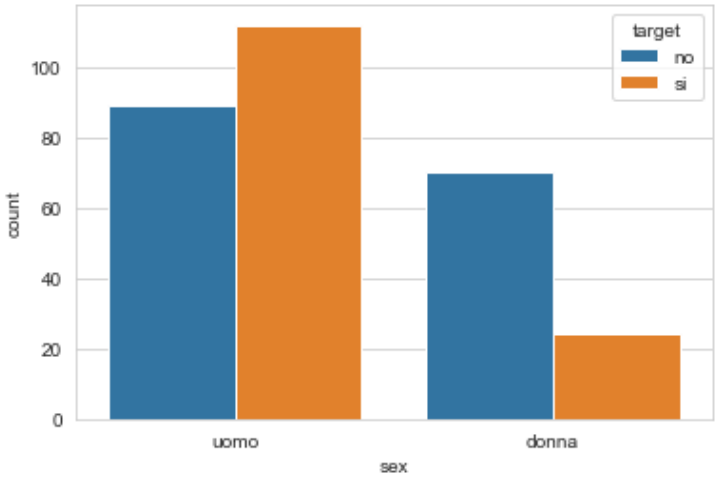




la colonna relativa agli anziani dice esattamente il contrario. Analizzando però la situazione tramite countplot, vediamo come gli anziani siano in numero molto limitato e probabilmente i presenti (sui 300 pazienti) non avevano una malattia, *dato* però che non possiamo considerare rilevante a causa lo scarso numero.

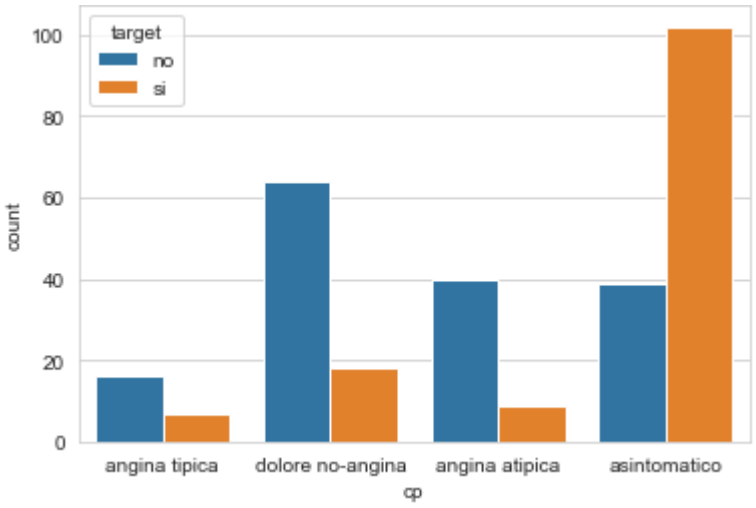
Sesso:

Nonostante la disparità dei numeri a livello di casi presenti nel dataset, si può notare che il sesso è un fattore di rischio infatti, in proporzione, gli uomini sono più tendenti a presentare malattie cardiache.



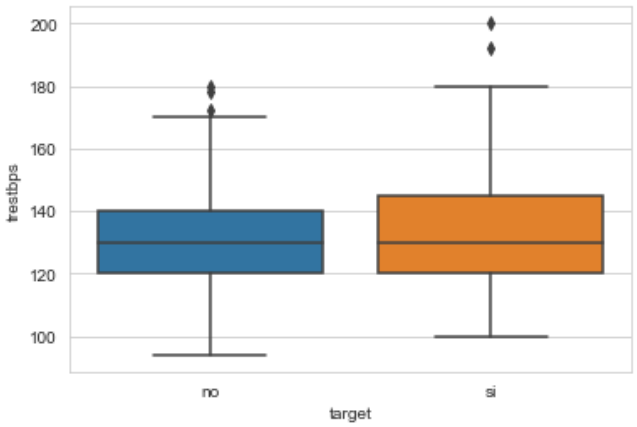
Dolore al petto:

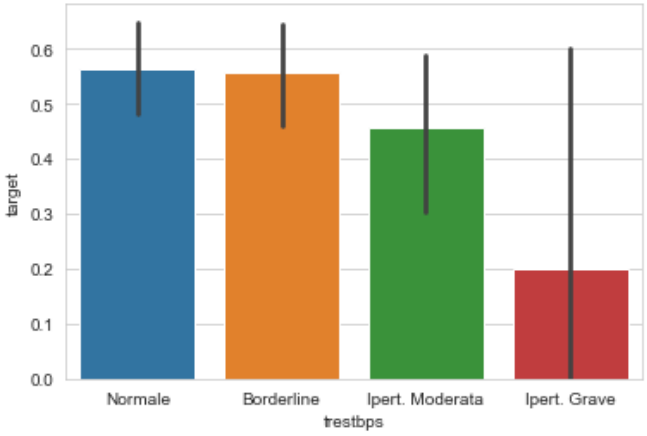
Considerando il risultato dell'analisi, non è molto chiaro se questo sia un fattore di rischio oppure no, dato che gli asintomatici presentano un rapporto molto più elevato di malattia rispetto al resto dei valori.



Pressione del sangue a riposo:

Dato il valore numerico ho deciso di esaminare dapprima il boxplot ma, a causa delle distribuzioni molto simili, ho deciso di analizzare anche il barplot con una suddivisione preventiva in 4 categorie con valori soglia che determinano la gravità del valore misurato…

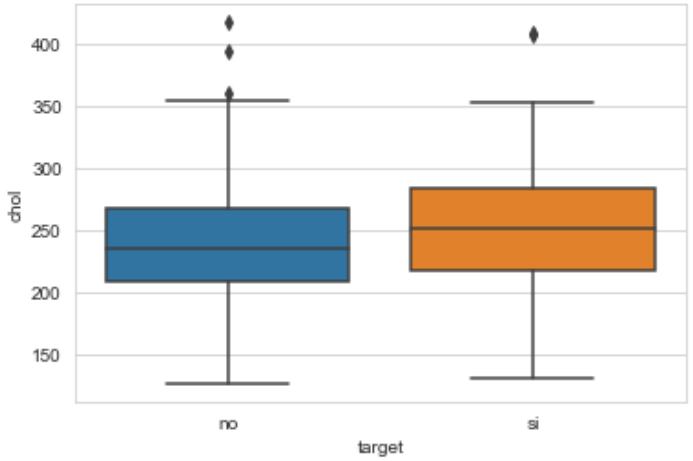




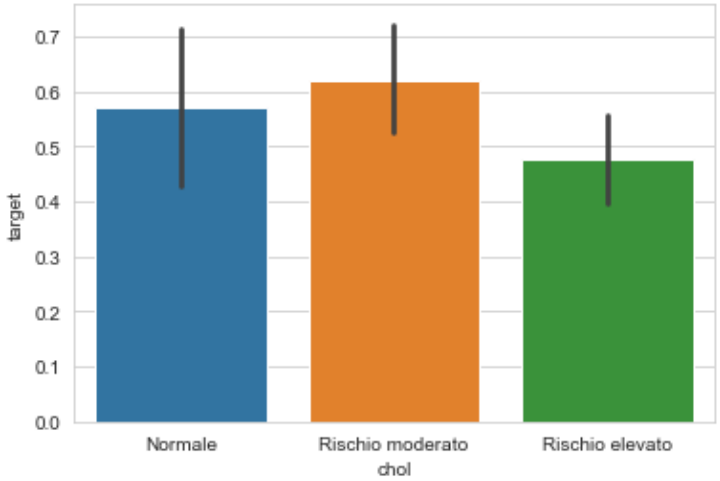
…dal quale possiamo constatare che, sopra una certa soglia, all'aumentare della pressione corrisponde un rischio più alto di avere una malattia cardiaca.

Colesterolo nel siero:

Come per la pressione del sangue, ho eseguito prima un boxplot e poi un barplot con una suddivisione in 3 categorie (sempre con valori di soglia per suddividere le tipologie).



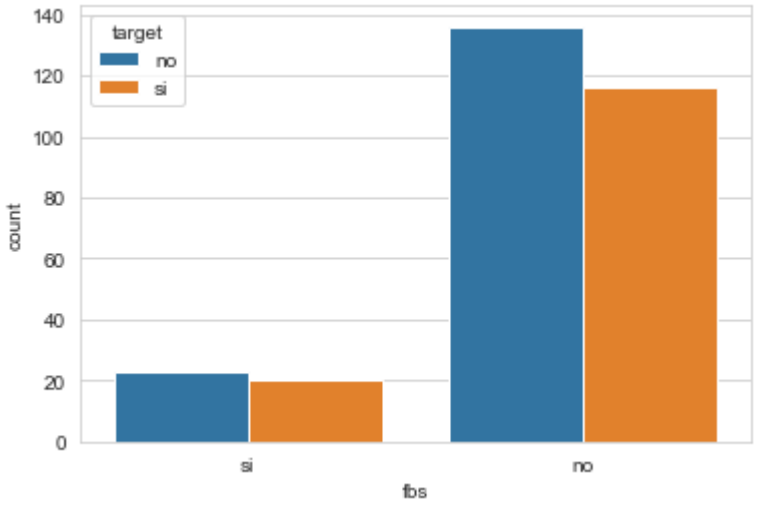
Purtroppo, come nel caso precedente, a parte qualche outlier non emerge una spiccata differenza tra i due casi…



…differenza che non viene risolta nemmeno da questa analisi, che denota una tendenza non radicale nel caso di "rischio elevato" al presentare una malattia del cuore.

Glicemia a digiuno (se > 120 mg/dl):

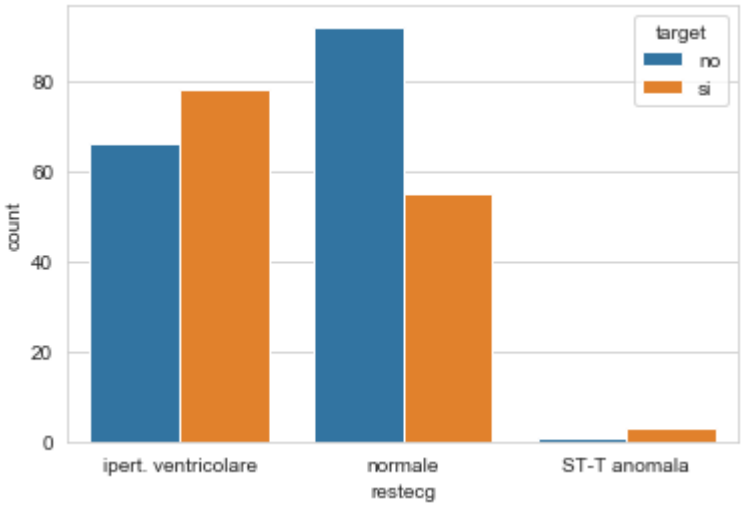
Escludendo il fatto che ci siano molti più pazienti senza questo valore positivo, si può notare come non sia un fattore correlato in qualche modo con la malattia cardiaca.



Elettrocardiogramma a riposo:

Quando un paziente avverte dolore da angina durante esercizio fisico, è un sintomo che fa pensare alla presenza di una malattia cardiaca; quando questo dolore è presente anche quando il paziente è a riposo, potrebbe significare una gravità della malattia non indifferente.

Questo potrebbe essere il motivo per cui ci sono pochi pazienti che presentano un'anomalia della linea ST, e quei pochi presentano quasi tutti una malattia.

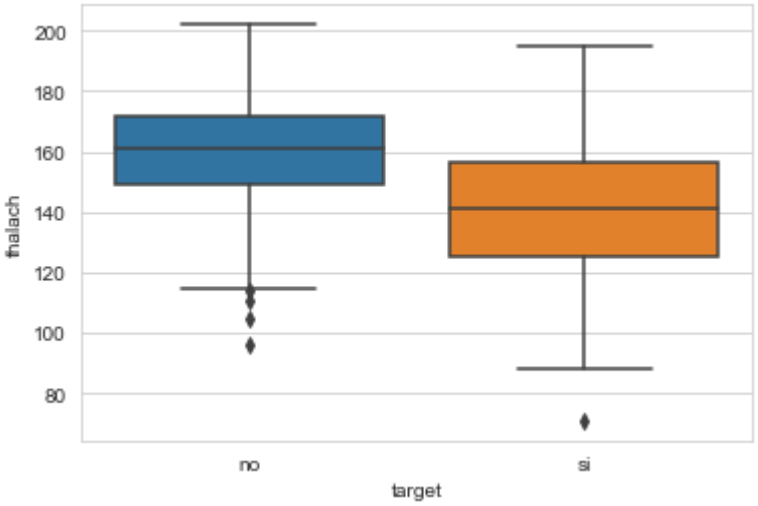


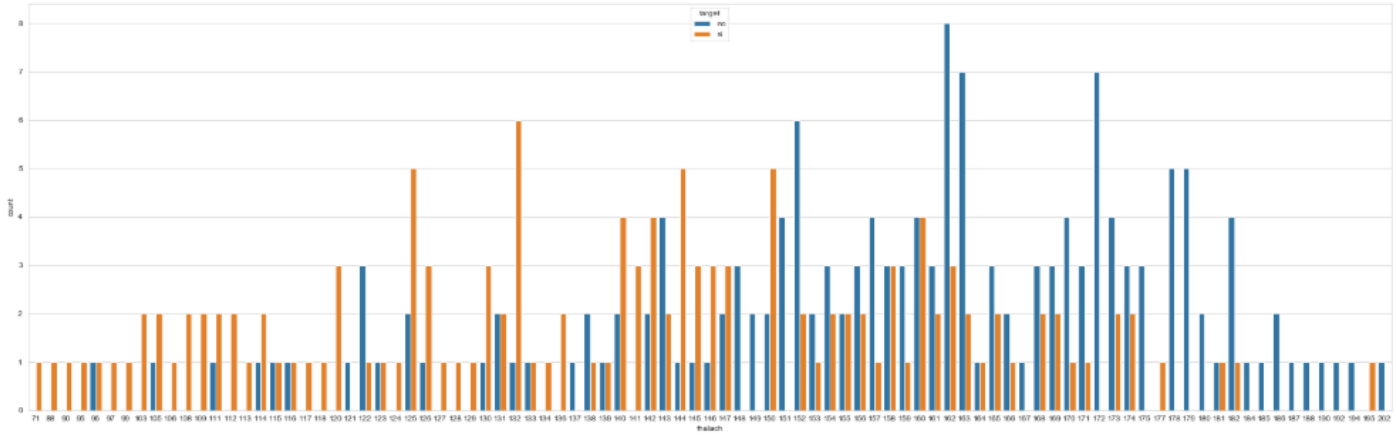
Per quanto riguarda una PROBABILE ipertrofia ventricolare, non sembra indicare nettamente la presenza della malattia (questo può essere spiegato dalla probabilità e non dalla certezza dell'ipertrofia).

Massimo battito cardiaco raggiunto:

analizzando il boxplot si può notare che la distribuzione dei valori associati a persone affette da malattie cardiache tende a valori inferiori rispetto alla controparte; data l'origine strana del risultato, ho voluto analizzare il countplot nella sua completezza con la conferma di ciò che era uscito dal boxplot.

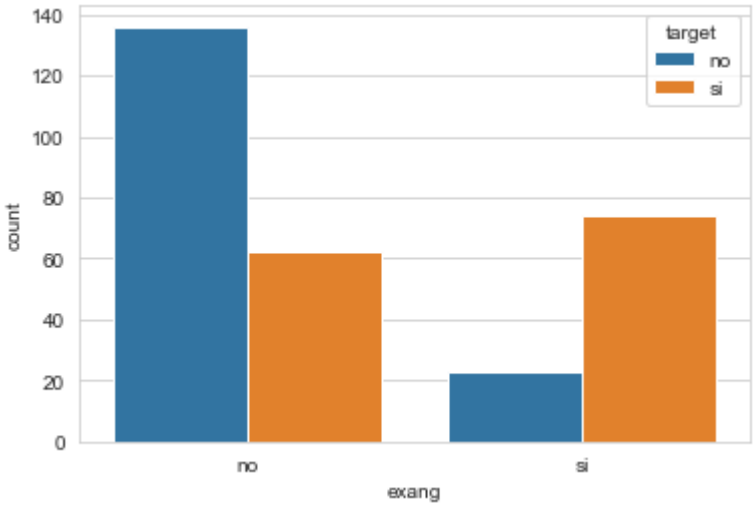
Andando a fare qualche ricerca sull'argomento, ho scoperto che un cuore "malato" fa fatica a raggiungere battiti elevati (un battito salutare non è costante, ma è dato da 220 - età).





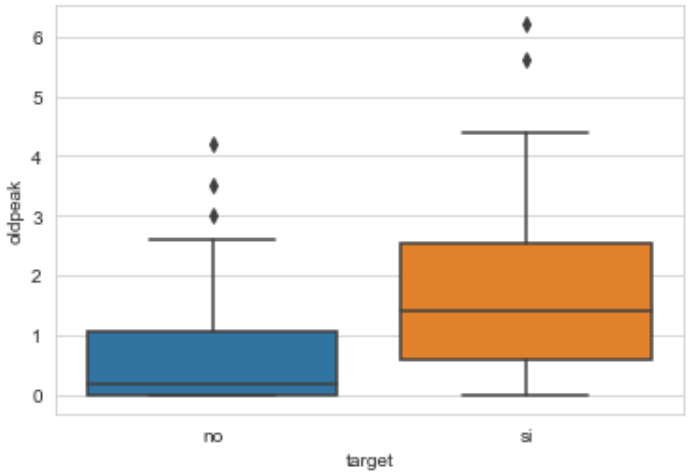
Angina sotto sforzo:

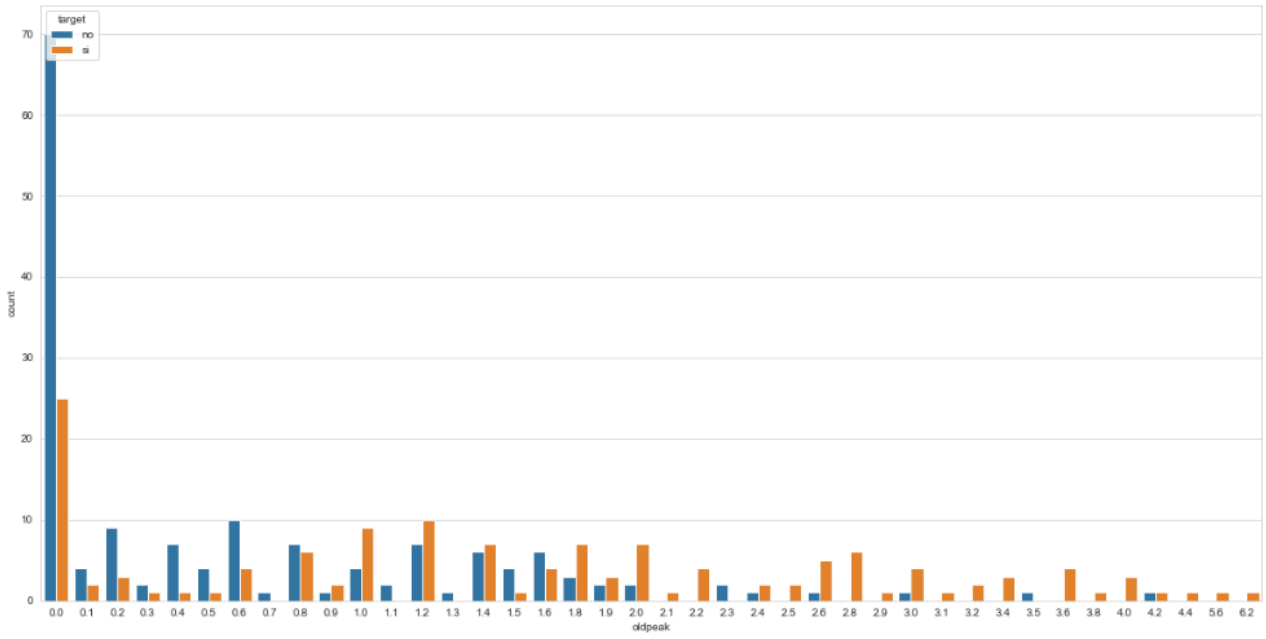
Dai dati raccolti risulta un fattore di rischio non indifferente, considerando che più del 70% dei pazienti che presentano angina presentano anche una malattia cardiaca.



Depressione del segmento ST indotta da esercizio rispetto ad una situazione di riposo:

Dal boxplot si nota una presenza della malattia all'aumentare della pendenza del segmento, tesi avvalorata dal countplot che mostra una quasi totale assenza di casi sani all'aumentare dell'indice di pendenza.

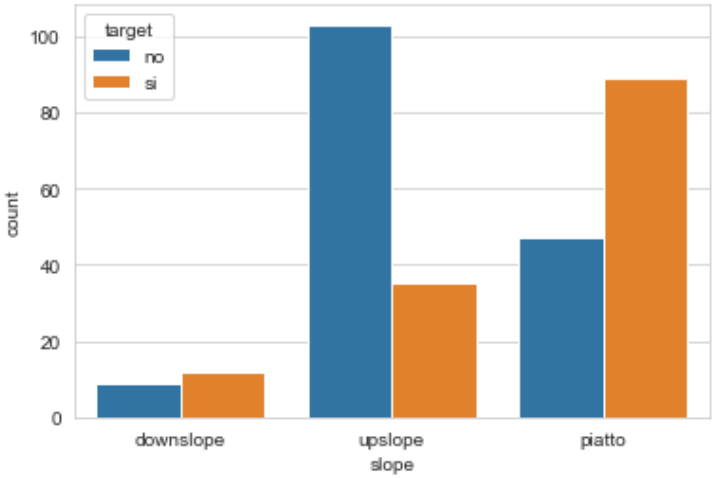


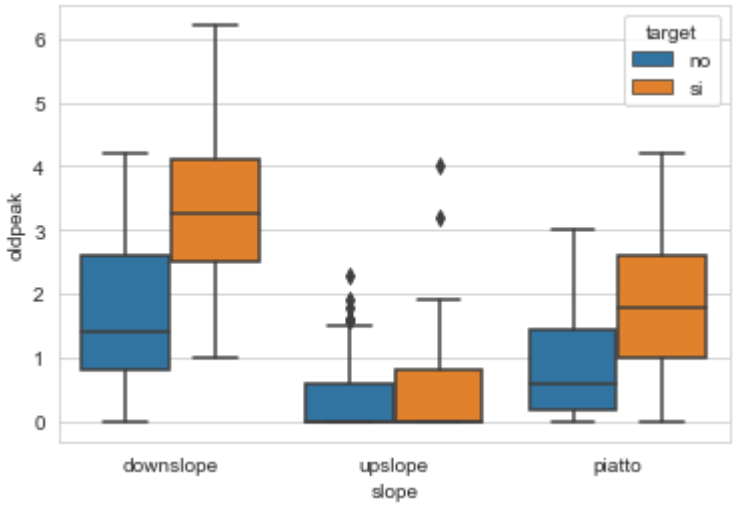


Pendenza del segmento ST durante esercizio:

In questo grafico i fattori determinanti derivano da una pendenza ascendente, che denota per lo più un'assenza di malattia, e un segmento ST che rimane piatto che invece comunica il risultato opposto.

Non è chiara invece la situazione per quanto riguarda una pendenza discendente: per questo ho unito il dato relativo alla depressione con la pendenza…

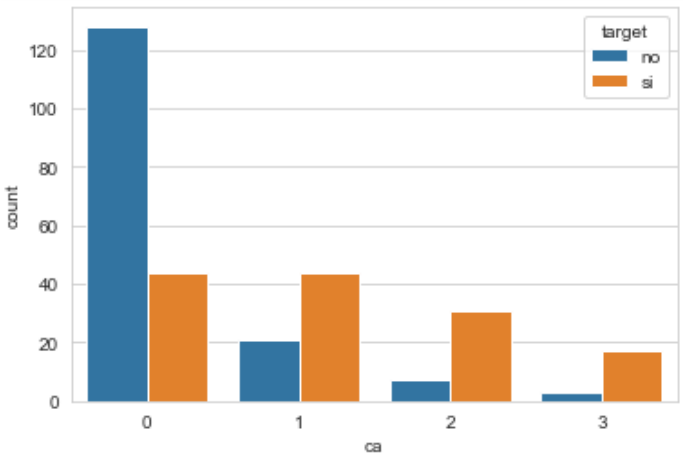




…dal quale si può chiaramente vedere come la depressione del segmento ST è il fattore di rischio maggiore all'aumentare del valore di questa.

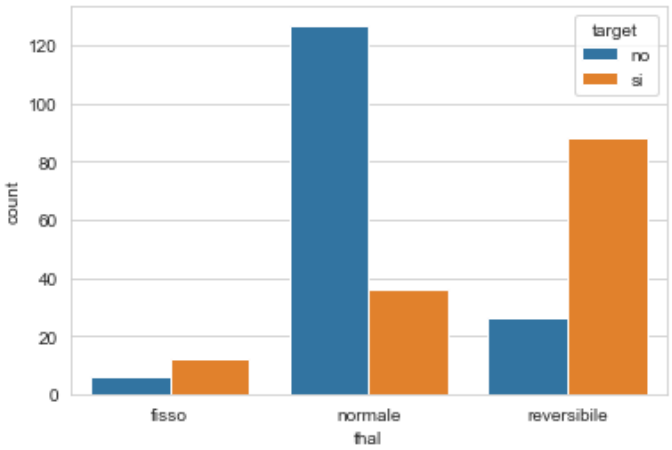
Numero di vasi principali colorati da fluoroscopia:

Questa features indica il numero di vasi RISTRETTI, da qui il motivo per cui più alto è il numero maggiore è la probabilità di avere una malattia cardiaca.

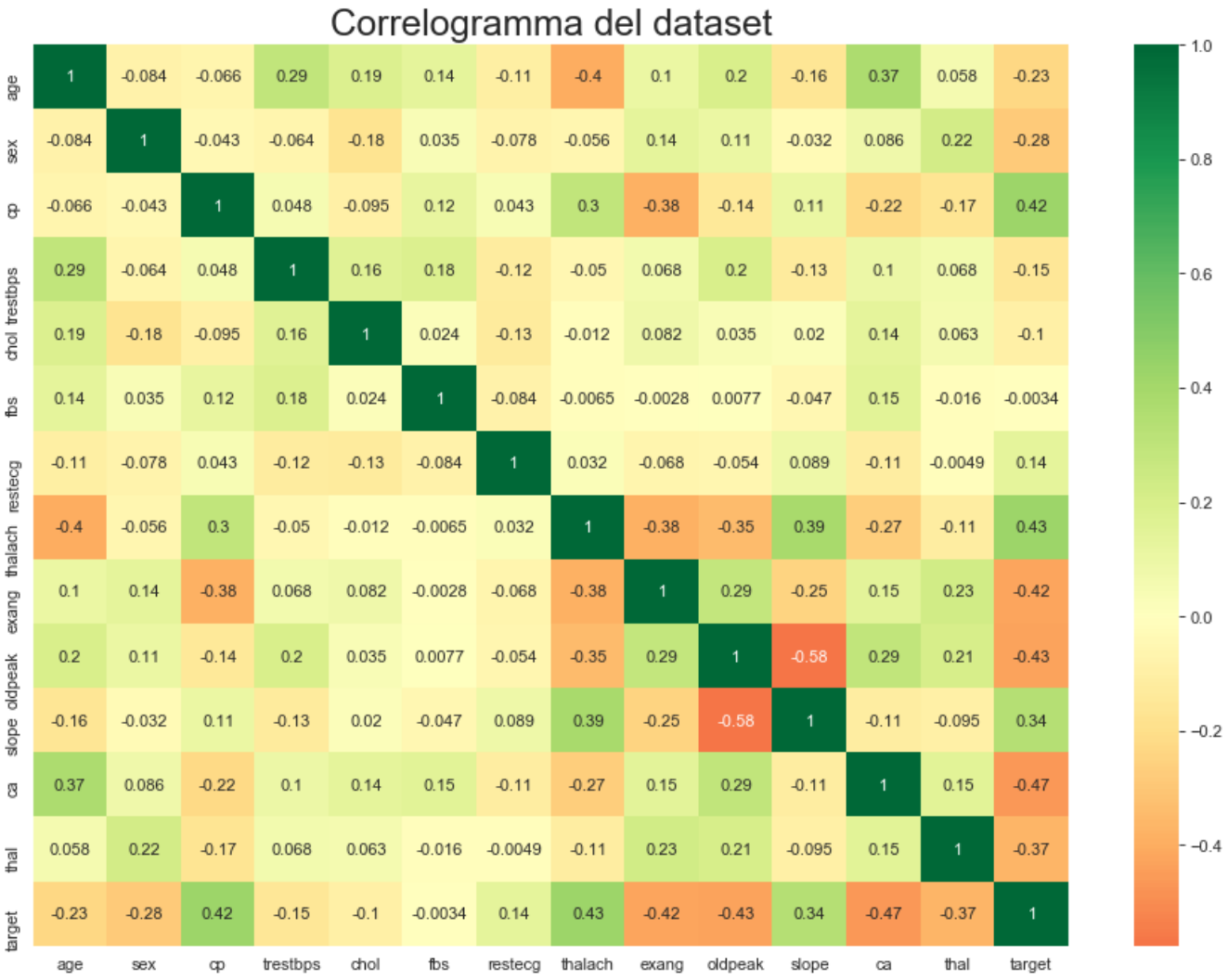


Difetto cardiaco:

Nonostante questo esame risulti parecchio invasivo per il paziente, possiamo considerarlo un fattore importante dato che nel caso di difetto la presenza di pazienti che mostrano una malattia è rilevante rispetto a quelli che non ce l'hanno.



Correlogramma dei vari attributi:



La matrice di correlazione mostra il "legame" che intercorre tra le varie features. Le correlazioni significative che possiamo notare sono tra:

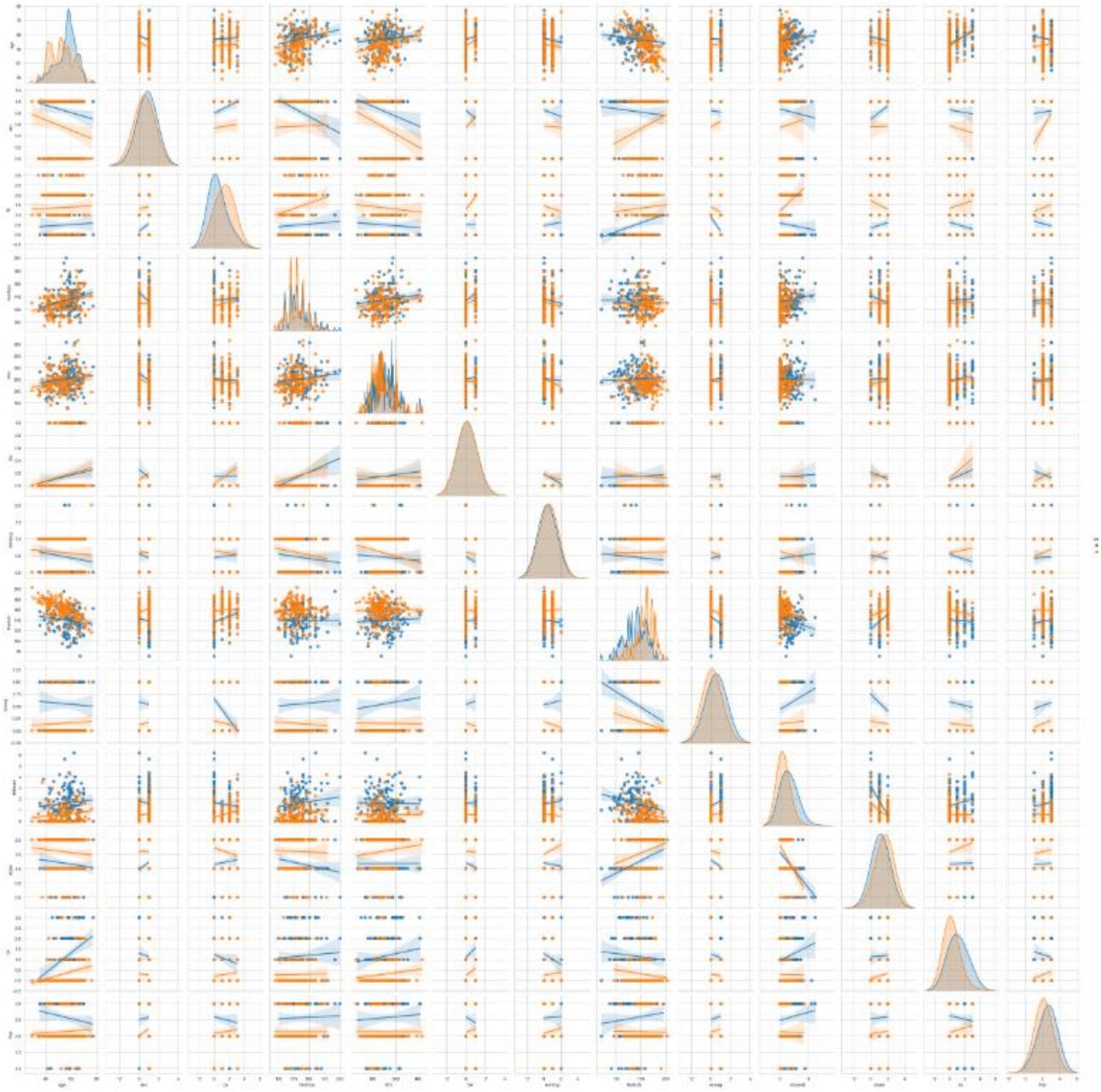
- età(age):
  - il numero di vasi colorati(ca) [positiva]
  - il massimo battito cardiaco raggiunto(thalach) [negativa]

- dolore al petto(cp)
  - angina sotto sforzo(exang) [negativa]

- massimo battito(thalach)
  - angina sotto sforzo(exang)[negativa]
  - depressione del segmento ST(oldpeak)[negativa]
  - pendenza del segmento ST(slope)[positiva]

- Depressione ST(oldpeak)
  - pendenza ST(slope)[negativa]

Le correlazioni venute alla luce dal correlogramma esaminato in precedenza possono essere analizzate nei grafici visibili nel pairplot che fa riferimento a tutte le features presenti nel dataset (esattamente come il correlogramma) .



Estrazione delle features

L'obiettivo delle tecniche utilizzate in questo paragrafo è ridurre la dimensionalità del dataset che si sta analizzando, rendendolo quindi più "leggero" in termini computazionali pur mantenendo le informazioni che ne permettono una buona analisi.

Le tecniche che sono state utilizzate sono la PCA (Principal Component Analysis) e la LDA (Linear Discriminant Analysis).

Sia la PCA che la LDA sono tecniche di trasformazione lineari ma, mentre la PCA predilige la direzione (componenti principali) che massimizza la varianza dei dati, la LDA ha come obiettivo il massimizzare la separazione (o la discriminazione) tra le differenti classi presenti che potrebbe risultare più utile in un contesto categorico (la PCA ignora le label associate ai dati).

PCA

Concetti teorici:

Algoritmo di apprendimento non supervisionato che ha come obiettivo l'individuazione delle direzioni in cui i dati presentano la massima varianza per la riduzione della dimensionalità del dataset d-dimensionale proiettandolo in un sottospazio k-dimensionale (con $k < d$).

Detto questo, è lecito chiedersi quale sia un valore di k che permetta di mantenere una buona mole di informazioni. Per scoprirlo si utilizza la coppia autovettore (i componenti principali) autovalore per determinare quale scartare per perdere meno informazioni possibili.

Approccio:

1. Standardizzazione dei dati (deviazione standard = 1, media = 0)
2. Calcolo degli autovalori e degli autovettori dalla matrice delle covarianze
3. Disposizione decrescente degli autovalori e selezione dei k autovettori (associati agli autovalori maggiori) dove k sarà la dimensionalità del nuovo sottospazio
4. Costruzione della projection matrix W (dagli autovettori selezionati)
5. Trasformazione del dataset originale tramite la matrice W per ottenere il nuovo sottospazio

Descrizione step:

1. Molto importante, se non fondamentale nella PCA, è la standardizzazione dei dati: spesso succede che i dati che abbiamo a disposizione siano caratterizzati da unità di misura non paragonabili o da un'ampiezza del sotto campione molto diversa. Questo porterebbe (nel caso di non standardizzazione) ad una valutazione errata da parte della PCA per il calcolo dei nuovi assi dato che si basa sulla deviazione standard di ogni variabile.

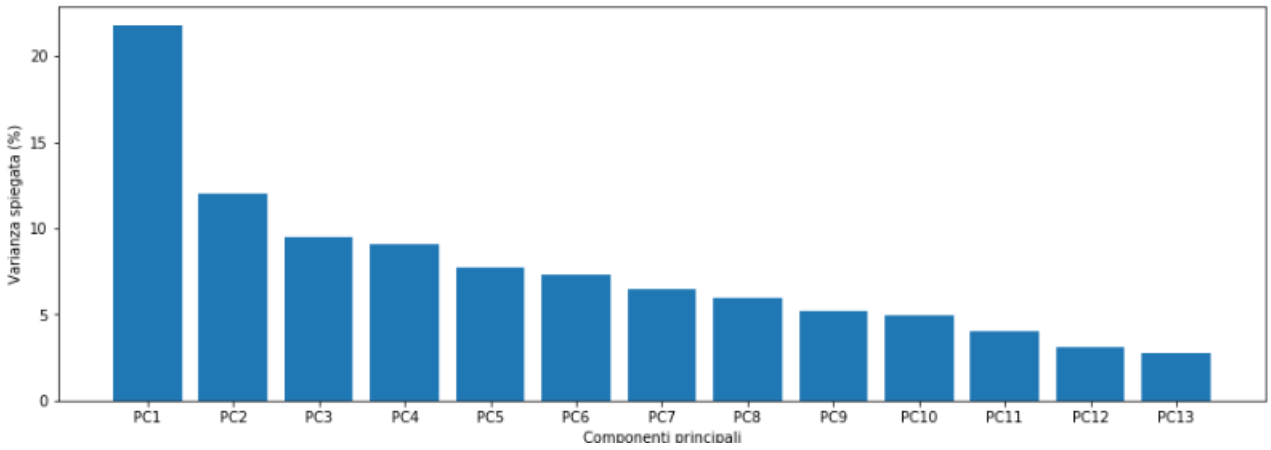
2. Dopo la standardizzazione, dalla matrice di covarianza estraggo gli autovettori che mi rappresentano le direzioni di massima varianza insieme agli autovalori che esprimono numericamente quanta varianza è presente in quella direzione.

   La matrice delle covarianze è una matrice *d x d* dove ogni elemento mi rappresenta la covarianza tra due features calcolata come:

$$\sigma_{jk} = \frac{1}{n-1} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

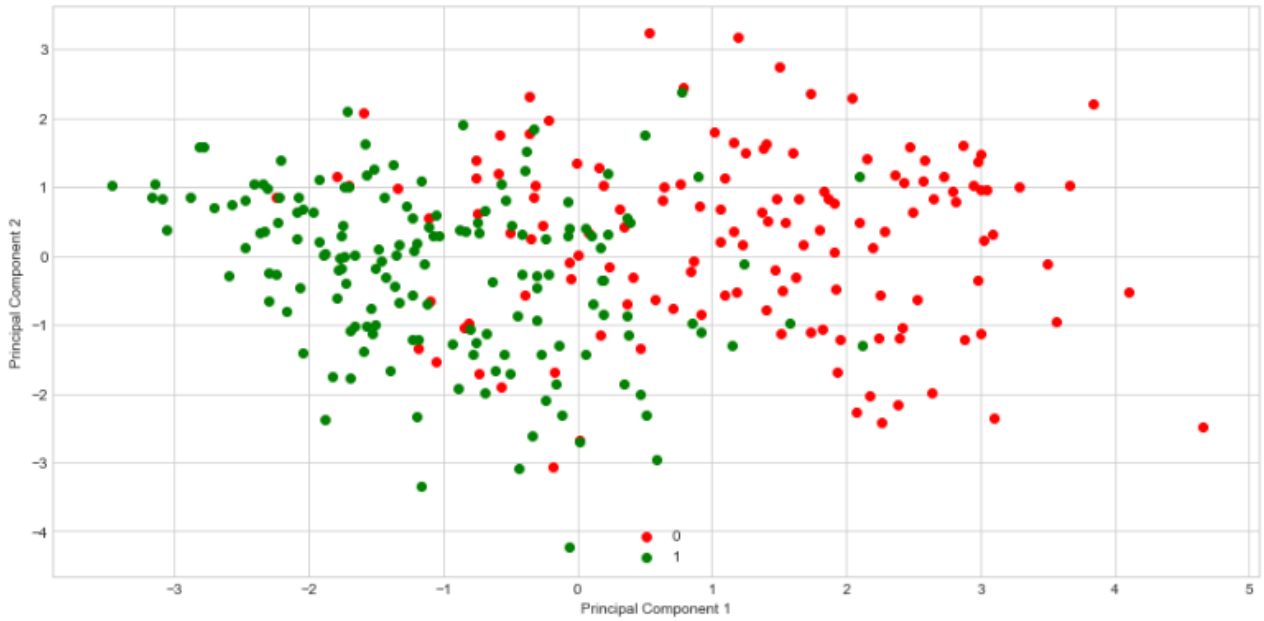
   Dove $\bar{x}$ è il vettore media d-dimensionale dove ogni valore rappresenta la media dei valori della colonna di una feature.

3. In modo da decidere quali autovettori tenere e quali scartare senza perdere troppa informazione, controllo gli autovalori e li ordino in ordine decrescente selezionando poi i "migliori" *k* autovettori associati a questi.
   Nel nostro caso, il grafico mostra chiaramente come, su 13 componenti, il primo da solo spieghi più del 20% della varianza totale e il secondo più del 10 %, mentre i restanti non vanno oltre al 10%.



4. La projection matrix non è altro che la matrice composta dai k autovettori che abbiamo selezionato come contenenti le maggiori informazioni.

5. In questo ultimo step si utilizza la matrice per proiettare il dataset nel nuovo sottospazio k-dimensionale (grafico nella pagina successiva)



Nonostante la poca varianza spiegata dal primo componente principale unito al secondo (circa 34%) si può notare come la separazione delle due classi sia già apprezzabile con un margine di errore evidenziato dall'area compresa tra -1 e 1.

LDA

Concetti teorici:

La Linear Discriminant Analysis non è altro che l'estensione supervisionata della PCA: l'obiettivo è quello di generare le componenti principali (come nella PCA) ma aggiungendoci le informazioni delle classi di appartenenza dei dati, in modo da massimizzare il distanziamento tra le classi (scatter between classes) e minimizzare il distanziamento all'interno della classe stessa (scatter within classes).

Approccio:

1. Calcolo:
   - La matrice della dispersione dei dati tra le classi (between scatter matrix)
   - La matrice della dispersione dei dati nelle classi (within scatter matrix)
2. Calcolo gli autovettori e i rispettivi autovalori per le matrici di dispersione
3. Ordino gli autovalori e seleziono i top k
4. Creo una nuova matrice contenente gli autovettori che mappano i k autovalori
5. Ottengo gli LDA components calcolando il prodotto scalare tra i dati e la matrice appena ottenuta

Descrizione step:

1. [A differenza della PCA, non c'è necessità di standardizzare i dati dato che non cambierebbe la loro classificazione]
   Calcolo la matrice della dispersione interna alle classi usando la formula:

   $\boldsymbol{S_w = \sum_{i=1}^{c} S_i}$ dove ***c*** è il numero totale di classi DISTINTE presenti nel dataset mentre

   $\boldsymbol{S_i = \sum_{x \in D_i}^{n} (x - m_i)\,(x - m_i)^T}$

   dove $\boldsymbol{m_i = \frac{1}{n_i} \sum_{x \in D_i} x}$ rappresenta la media della classe i_{esima}.

   (**x** è un sample e ***n*** è il totale dei sample presenti in una classe)

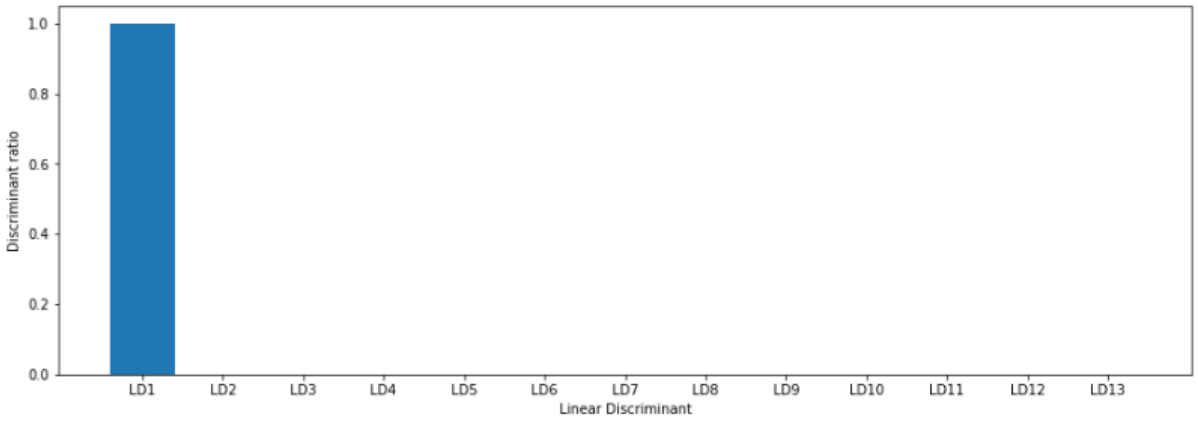
   Dopodiché calcolo la matrice della dispersione tra le varie classi tramite la formula:

   $\boldsymbol{S_B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T}$

   dove $\boldsymbol{m_i = \frac{1}{n_i} \sum_{x \in D_i} x}$ rappresenta la media della classe i_{esima}.

2. Per calcolare autovalori e autovettori risolvo il problema generico $\boldsymbol{S_W^{-1} S_B}$

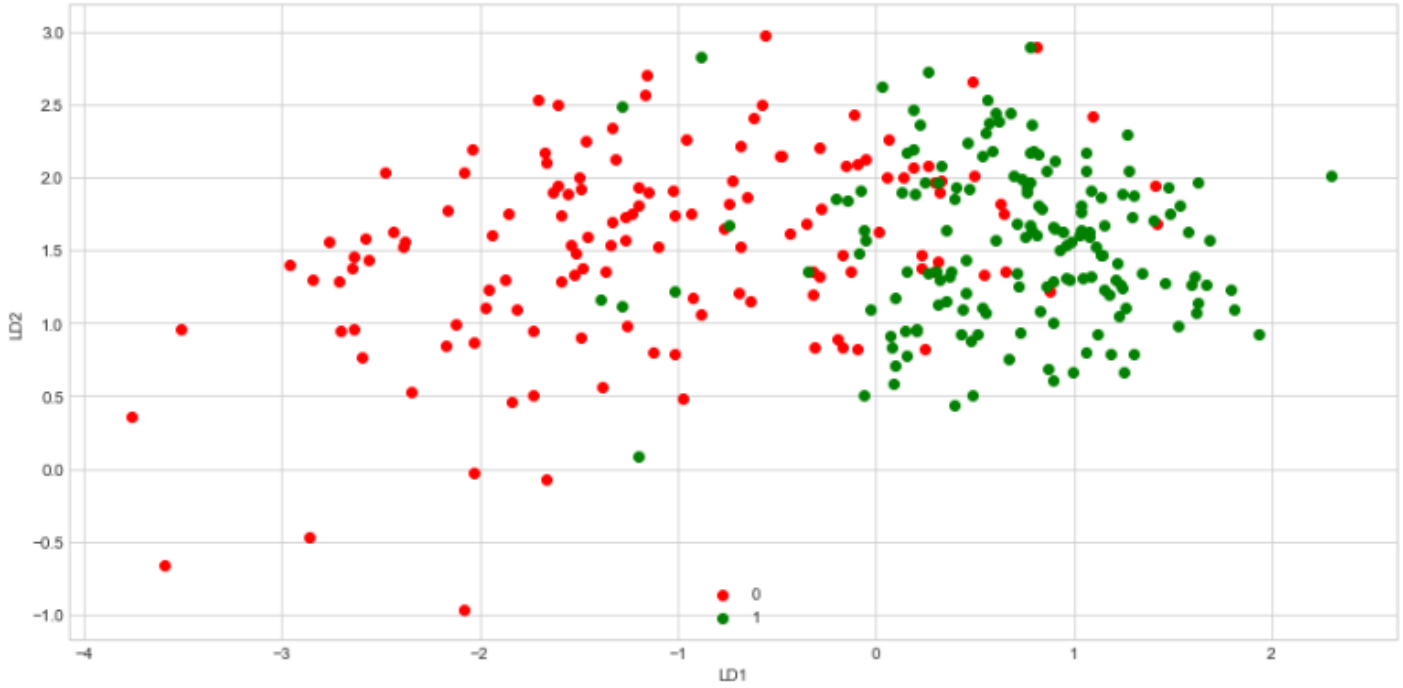
3. Ordino gli autovalori dal maggiore al minore e seleziono i primi k autovettori corrispondenti.
   Nel nostro caso possiamo notare un risultato particolare: la quasi totalità della varianza è spiegata dal primo componente principale, rendendo quasi inutile considerare gli altri 12.



4. Dopo aver selezionato i k autovettori che spiegano la maggior varianza, creo una nuova matrice **W** con questi al suo interno.

5. Come ultimo passaggio, creo il nuovo spazio k-dimensionale applicando il prodotto scalare tra i dati e la nuova matrice ottenuta $\boldsymbol{Y = X \cdot W}$
   Dove:
   - ***X*** è una matrice ***n*** x ***d*** (con ***n*** samples e ***d*** = numero di dimensioni)
   - ***Y*** è una matrice ***n*** x ***k*** con ***n*** samples e ***k*** dimensioni (***k*** < ***d***)

Anche se nel nostro caso la varianza è spiegata quasi totalmente dal primo componente, ho voluto comunque utilizzare i primi due (che insieme formano una *class-discriminability*) per avere un grafico in 2 dimensioni, a mio parere migliore a livello di comprensione.



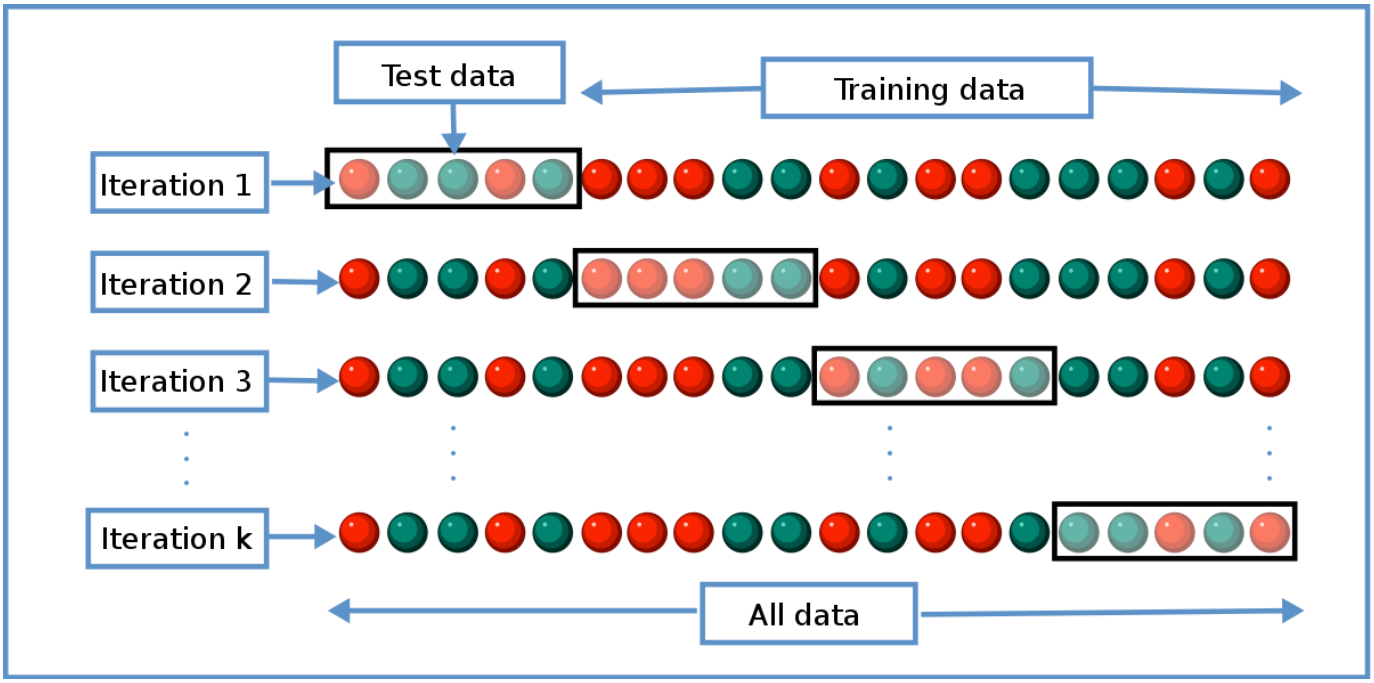
La separazione ottenuta è molto più evidente rispetto a quella ottenuta plottando i primi due componenti della PCA, anche se si sapeva già guardando i numeri.

Tecniche di classificazione

Terminata l'analisi sul dataset, ho voluto applicare tre diverse tecniche di classificazione per determinare:

- quanto il dataset potesse essere utile per predire lo stato del paziente (malattia presente/assente);
- quale tecnica sia più consona allo scopo, paragonando i risultati.

Per testare i migliori parametri per le varie tecniche ho utilizzato la *k-fold cross-validation*: questa si basa sull'idea di dividere randomicamente il dataset in k sottoparti uguali, dopodiché si lascia fuori una delle parti che verrà utilizzata come test mentre il modello si allena sulle restanti (a ciclo si ripete per tutte le parti). Dalla media dei valori predittivi ottenuti nei k esperimenti si calcola la media cumulativa finale che misura l'accuratezza del modello.



Le tecniche di classificazione che ho scelto sono:

- Logistic regression
- Decision Tree
- K-nearest neighbors

Logistic regression

La regressione logistica è un metodo statistico per predire classi binarie: il target è dicotomico (ovvero che sono possibili solo due risultati) per definizione.

La regressione logistica è un caso particolare della regressione lineare dove il target è di origine categorica: questo è reso possibile attraverso l'uso di una funzione logit.

Funzionamento

In questi modelli si utilizza la funzione logistica $\boldsymbol{P(Y=1) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}$.

Manipolando questa funzione si ottiene $\boldsymbol{\frac{P(Y=1)}{1-P(Y=1)} = e^{\beta_0+\beta_1 X}}$, dove la quantità $\boldsymbol{\frac{P(Y=1)}{1-P(Y=1)}}$ è definita odds e restituisce un valore compreso tra [0,1] se in input riceve un valore compreso tra [0, +∞].

Applicando il logaritmo naturale all'odds otteniamo $\boldsymbol{\ln \frac{P(Y=1)}{1-P(Y=1)} = e^{\beta_0+\beta_1 X}}$ che prende il nome di logit ed è lineare in X.

Possiamo dedurre che, nella regressione logistica, $\boldsymbol{\beta_1}$ risulta legato alla variazione del logit e non a quella cui è legato non linearmente.

La regressione logistica, a livello di codice, ha alcuni parametri di tuning tra cui:

- **C**: parametro regolatore inverso ***C = 1/λ***. Diminuendo C il valore di lambda aumenta, usato per regolare e contrastare il fenomeno dell'overfitting. Determina il grado di libertà.
- **Penalty**: usata per specificare la norma usata nella penalizzazione.
- **Class weight**: se specificato, permette l'utilizzo di un distributore automatico di pesi basato sulla proporzione inversa alla frequenza delle classi.

Vantaggi/svantaggi

I vantaggi di questa tecnica sono radicati nella sua semplicità: non richiede grandi capacità di calcolo, è di semplice implementazione e interpretazione e non richiede lo scaling delle features.

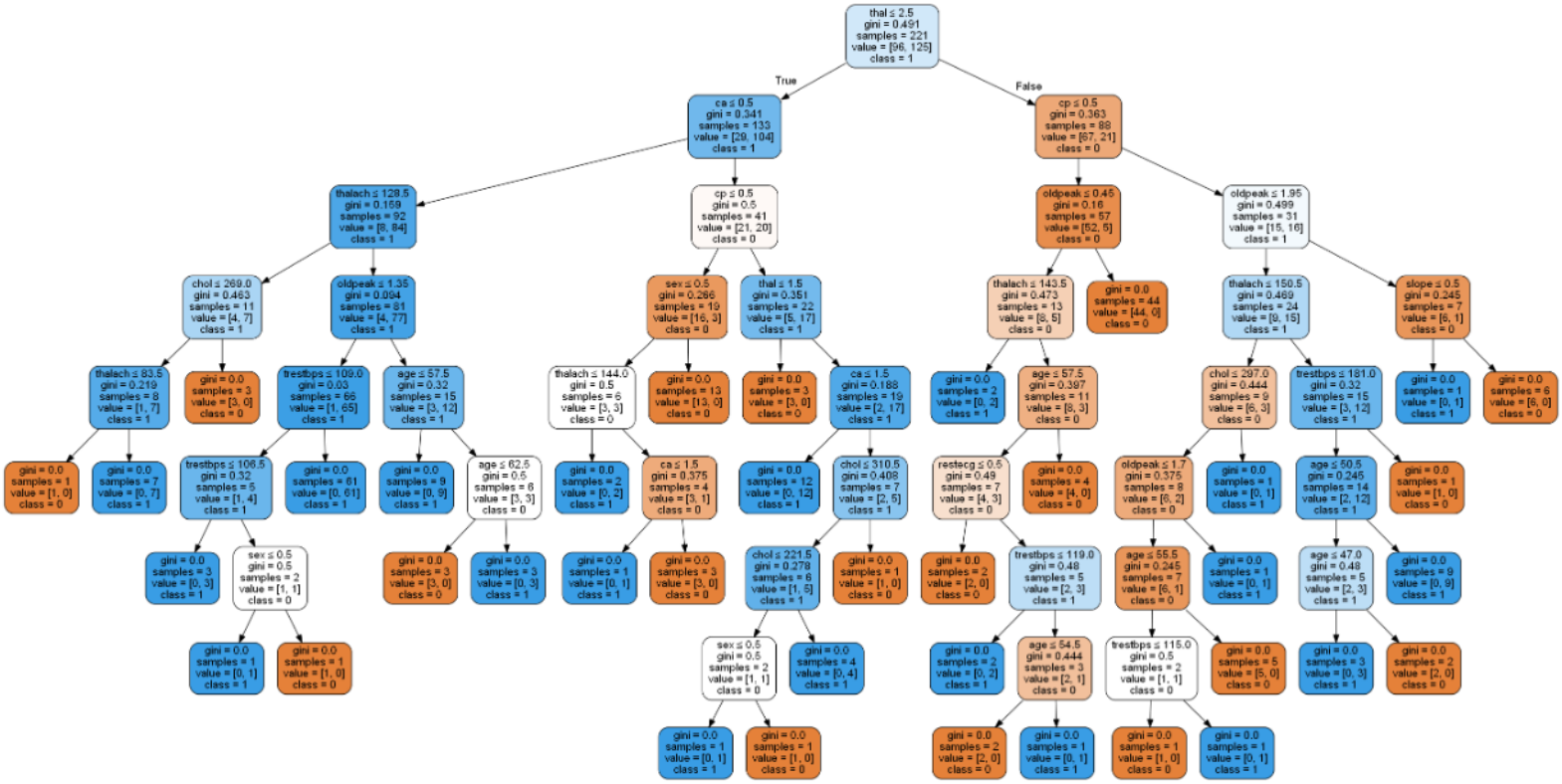
I principali vantaggi invece risiedono nell'impossibilità di coprire un vasto numero di features categoriche ed è vulnerabile all'overfitting.

Decision Tree

Un decision tree è una struttura ad albero (simile ad un flowchart) dove ogni nodo interno rappresenta una feature, i rami rappresentano le regole di decisione mentre le foglie mostrano il risultato.

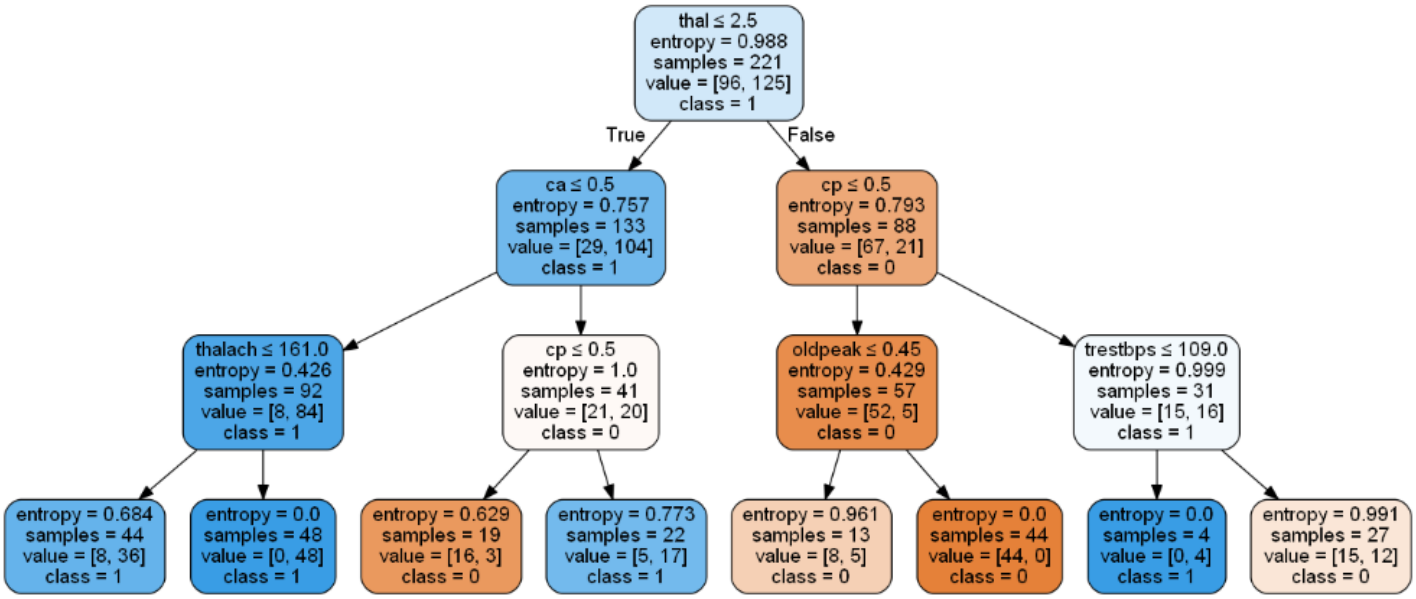
Il nodo in cima è conosciuto come radice e impara a separarsi in base al valore dell'attributo che rappresenta; l'albero viene così ripartito ricorsivamente fino ad arrivare ad una profondità limite imposta da noi o dalla terminazione delle features.

A differenza di altri metodi di ML, il decision tree può essere inserito nella categoria dei "white box algorithm" ovvero algoritmi di apprendimento che mostrano all'utente come hanno appreso e portato a termine determinate scelte (questo è uno degli aspetti più interessanti): data questa particolare caratteristica, possiamo analizzare le feature ritenute più segnanti dal metodo per quanto riguarda il nostro caso plottando l'albero decisionale risultante:



Non avendo inserito nessun limite, l'algoritmo va avanti fino a che non terminano gli attributi rendendo la lettura abbastanza complessa.

Per questo motivo si può "ottimizzare" il processo utilizzando il *pre-pruning*, una tecnica di early stopping che interviene durante la creazione dell'albero; questa tecnica ha un duplice vantaggio, infatti se da una parte impedisce un'elevata specializzazione dell'algoritmo (quindi blocca il fenomeno dell'overfitting, se presente), dall'altra rende più comprensibile il risultato che si ottiene:



Funzionamento

L'idea alla base è la seguente:

1. Si selezione l'attributo/feature migliore tramite l'Attribute Selection Measures(ASM) per dividere i dati.
2. Si fa diventare l'attributo un nodo decisionale e si spezza il dataset in set più piccoli
3. Si ripete questo processo fino a che:
   - Tutte le tuple appartengono allo stesso attributo
   - Non vi sono più attributi rimanenti

Come si può facilmente intuire, una parte fondamentale dell'algoritmo è lo split dei dati: questo può avvenire attraverso vari termini, tra cui il Gini index e l'entropia:

- Gini index: $Gini(D) = 1 - \sum_{i=1}^{m} Pi^2$, dove ***Pi*** è la probabilità che una tupla in D appartenga alla classi Ci (raggiunge lo 0 quando appartiene ad una classe specifica);

- Entropia/information gain: data la definizione di entropia come "impurità in un gruppo di samples", viene calcolata la differenza tra l'entropia prima e dopo lo split nel caso dei vari attributi e viene scelto l'attributo con il valore di gain maggiore come nodo.

  Numericamente i due sono molto simili, portando a volte allo stesso risultato.

Vantaggi/svantaggi

Come vantaggio gli alberi decisionali hanno sicuramente la facile interpretazione del risultato, in aggiunta alla cattura di pattern non lineari e alla non assunzione della distribuzione data la non-parametrica natura dell'algoritmo.

Contro hanno una sensibilità particolare al rumore (si rischia l'overfit su dati errati) e al cambiamento ,anche non sconvolgente, dei dati che può portare alla definizione di un albero diverso (aggiustabile tramite bagging e boosting).

KNN (K-Nearest Neighbors)

Algoritmo non parametrico che si basa sul concetto di classificare un dato in input considerando la classe più "vicina" ad esso rilevata in fase di addestramento. Viene definito un algoritmo pigro, perché il training set È il modello, riducendo a zero il tempo destinato in altri algoritmi alla creazione ,appunto, del modello: questo però significa che tutta la forza computazionale viene sfruttata durante la fase di testing.

Funzionamento

K è il numero di vicini meno distanti dal dato che risulta essere la base decisionale dell'algoritmo (solitamente viene scelto un numero pari se le classi target sono 2).

Prima di tutto si identificano i k punti più vicini al punto che si vuole classificare dopodiché si applica l'etichetta valutando la maggioranza.

La "vicinanza" può essere calcolata tramite la distanza euclidea, la distanza Manhattan o altre formule che definiscano numericamente lo spazio presente tra i punti.

Non esiste un numero ottimale di K: se si scegli un numero troppo basso c'è pericolo che il rumore diventi un fattore rilevante, mentre numeri elevati rendono l'algoritmo computazionalmente dispendioso.

Curse of dimensionality

Questo algoritmo performa meglio con un numero limitato di dimensioni perché, all'aumentare delle features, il numero di dati richiesto per mantenere un buon livello di accuratezza generale cresce esponenzialmente.

Dato che nel nostro caso è presente un numero non indifferente di features, ho provato a far distinzione tra un KNN applicato al dataset così com'è e un KNN applicato ad un dataset con le features standardizzate: questo perché, essendo un algoritmo basato sulla distanza, ed essendo molto spesso questa "euclidea", la standardizzazione dei dati permette di eliminare l'influenza di una features rispetto ad un'altra basata sull'unità di misura, concentrando invece la decisione sull'effettivo impatto che questa ha sul target.

Questo cambiamento risulta largamente visibile anche nel nostro caso:

VALORE	DATASET ORIGINALE	STANDARDIZZAZIONE
ACCURATEZZA	0.6891	0.8243
PRECISIONE	0.6279	0.7692
RECALL	0.7941	0.8823

Le differenze nei risultati ottenuti sono palesi sotto ogni metrica misurata, convalidando l'ipotesi teorica fatta in precedenza.

Svantaggi

Il KNN rimane molto suscettibile alla presenza di outliers e al rumore; in più la scelta del giusto parametro K rimane determinante affinché l'algoritmo performi in modo corretto.

Classificatori a confronto

Confusion matrix

Concetto di base:

L'analisi tramite confusion matrix ci permette di confrontare il numero di campioni che non sono stati classificati correttamente con il numero di campioni per cui è stata predetta la classe corretta tramite questo schema:

	1 (Predicted)	0 (Predicted)
1 (Actual)	True Positive	False Negative
0 (Actual)	False Positive	True Negative

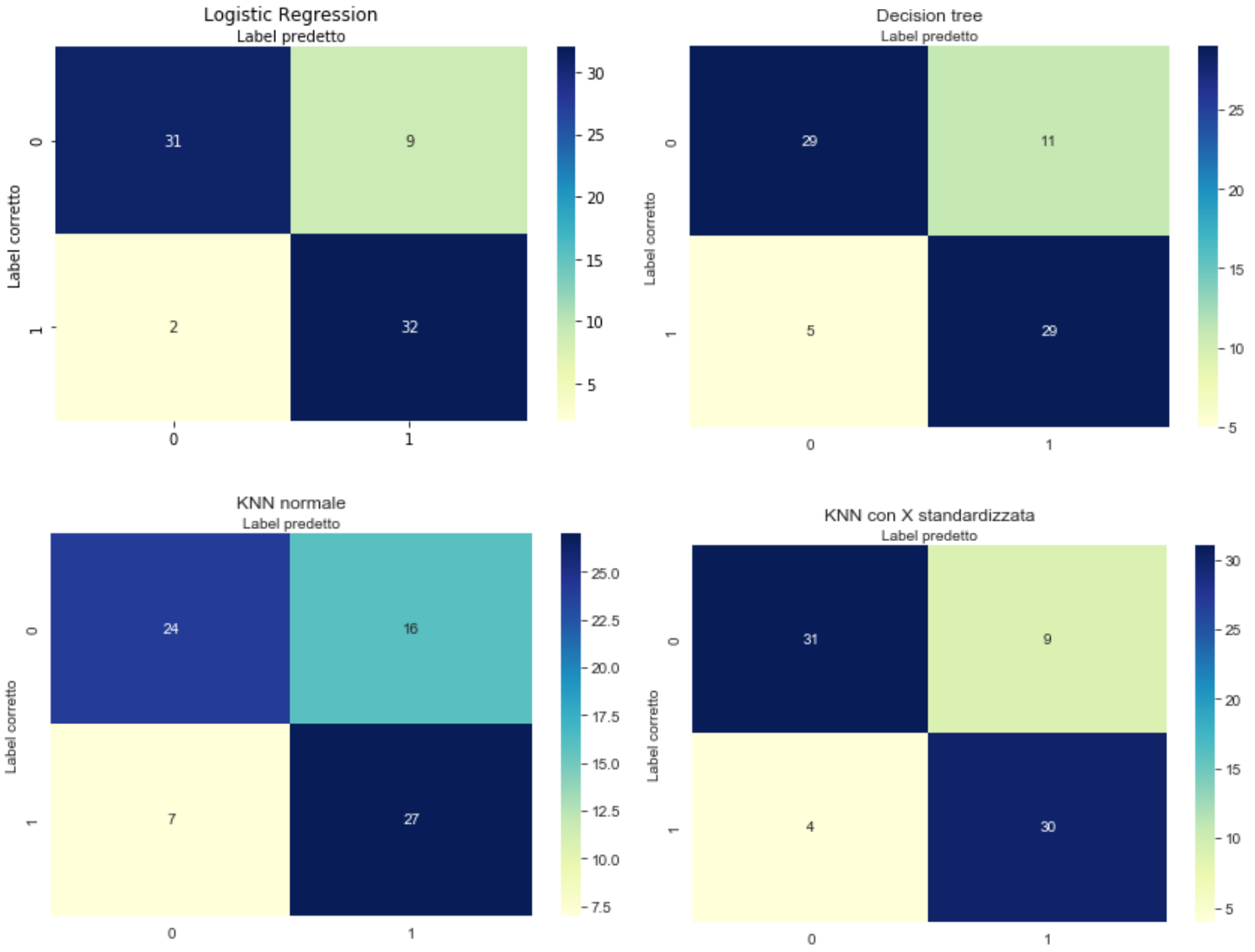
(la definizione di positivo/negativo è convenzionale, si utilizza come riferimento "positivo" la classe in cui siamo maggiormente interessati).

- True positive: casi positivi correttamente etichettati
- True negative: casi negativi correttamente etichettati
- False positive: casi negativi etichettati come positivi
- False negative: casi positivi etichettati come negativi

Dalla confusion matrix è inoltre possibile estrapolare due dati importanti per analizzare la bontà di un modello, ovvero:

- Recall $= \frac{TP}{TP+FP}$: questo dato indica quanti campioni positivi sono stati identificati
- Precision = $\frac{TP}{TP+FN}$: questo dato invece ci indica quanti campioni positivi sono stati identificati correttamente

Confusion matrix dei modelli applicati



Dalle confusion matrix messe a confronto diretto si può notare come, oltre al KNN applicato senza standardizzazione dei dati, gli altri modelli abbiamo delle performance generali simili, con giusto qualche elemento di scarto.

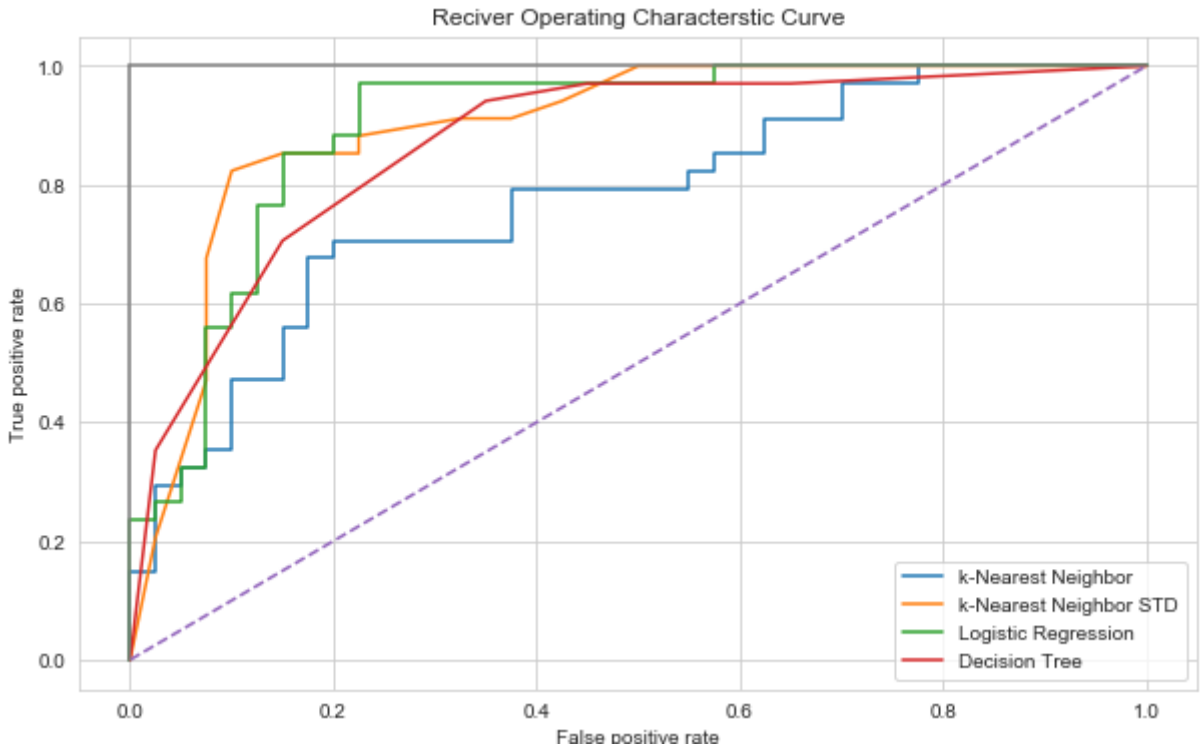
Come commento comune a tutti i modelli possiamo vedere come essi tendano più ad etichettare un paziente come “sano” (valore: 1) quando invece esso risulta “malato” (valore: 0) invece che l’opposto.

ROC curve (Receiver Operating Characteristic)

Concetto teorico

Altra analisi molto importante che si può analizzare attraverso il rapporto dei TP (True positive) e dei FP (False Positive) è sicuramente la curva ROC, creata tracciando il valore del True Positive Rate (frazione dei veri positivi, i.e. **sensibilità**) rispetto al False Positive Rate (frazione dei falsi positivi, i.e. **specificità**).

ROC dei modelli applicati



Come è visibile dal confronto delle confusion matrix precedente, il KNN sul dataset senza standardizzazione è il modello che tende di più a classificare falsi positivi; questo fenomeno è riconoscibile dalla crescita molto frastagliata e con segmenti molto più lunghi sull'asse del FPR rispetto agli altri modelli.

LDA applicata prima dei modelli

Preconcetti basati sulla teoria

Considerando che, se l' obiettivo principale è il classificare con la massima accuratezza possibile i pazienti futuri che si presenteranno in ospedale, allora possiamo sicuramente prendere come obiettivo secondario (ma altrettanto importante) l'efficienza computazionale dei metodi utilizzati. In questo modo, in caso di aumento dei record nel dataset (cosa che sarebbe solo gradita, considerando il numero ridotto di tuple utilizzato in questo caso), le performance sarebbero accettabili anche senza adottare un hardware più costoso.

Dalla riduzione della dimensionalità, un dato interessante è sicuramente venuto fuori dall'LDA che, tramite un unico componente, riesce a spiegare il 99.99% della varianza.

Questo significa che, tramite questa tecnica, potrei ridurre le dimensionalità del mio dataset da 13 a 1 senza perdere quasi nessuna informazione sulla varianza dei dati.

Il metodo che ne gioverebbe di più sarebbe sicuramente il KNN dato che soffre della curse of dimensionality e nel nostro dataset sono presenti poche entries.

Risultati ottenuti

Logistic Regression

VALORE	DATASET COMPLETO	LDA
ACCURATEZZA	0.8513	0.8648
PRECISIONE	0.7804	0.8372
RECALL	0.9411	0.9230

Confrontato con i risultati ottenuti con il dataset intero, la performance è ,anche se può sembrare pressoché invariata, migliorata: vi sono delle piccole percentuali di perdita nella recall, ma possiamo vedere come la precisione con cui vengono assegnate le classi di appartenenza migliori.

Decision Tree

VALORE	DATASET COMPLETO	LDA
ACCURATEZZA	0.7837	0.8378
PRECISIONE	0.7250	0.8648
RECALL	0.8529	0.8205

Anche in questo caso, grazie alla riduzione da parte della LDA, vediamo un aumento del 5% sull'accuratezza, portandoci ad un totale maggiore dell'80%, con un netto miglioramento nella precisione e qualche piccola perdita nella recall.

KNN

VALORE	DATASET COMPLETO	LDA
ACCURATEZZA	0.6891	0.8918
PRECISIONE	0.6279	0.8974
RECALL	0.7941	0.8974
N. NEIGHTBORS	14	12

A differenza degli altri classificatori, nel KNN possiamo notare come l'accuratezza abbia giovato dalla riduzione della dimensionalità in maniera significativa, si parla di più del 20% (la curse of dimensionality impatta molto dato che abbiamo pochi dati a disposizione) ma non solo: sia precisione che recall sono arrivate quasi al 90% mentre abbiamo ridotto il numero di vicini di 2.

Conclusioni

In generale si può affermare che l'obiettivo è stato raggiunto e anche con dei risultati tutt'altro che esigui: se si considera la dimensionalità del dataset (13 features escluso il target) e il numero limitato di pazienti presenti (circa 300, considerando che sono stati anche rimosse delle righe causa valori errati), il minor *classification rate* ottenuto applicando tutti i modelli (escludendo il KNN applicato a tutto il dataset, dato che si sapeva ancor prima dell'applicazione che avrebbe avuto scarsi risultati) è di:

- **78,37%** senza riduzione della dimensionalità
- **83.78%** con la riduzione della dimensionalità ottenuta tramite LDA

Mentre il maggiore:

- **85,13%** senza riduzione della dimensionalità
- **89.18%** con la riduzione della dimensionalità ottenuta tramite LDA

I metodi scelti e i modelli derivanti potranno essere utilizzati in ambito medico-ospedaliero per determinare, date le analisi utilizzate nel dataset usato come base, se un paziente rischia di avere una malattia cardiaca con una buona affidabilità.