

UNIVERSITÀ DEGLI STUDI DI PAVIA

FACOLTÀ DI INGEGNERIA

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

REPORT APPRENDIMENTO AUTOMATICO IN MEDICINA

OGGETTO: HEART_DISEASE

GIOELE MARUCCIA

MAT.452306

A.A 2017/2018

SOMMARIO

Definizione del problema	4
Obbiettivo	4
Goal dell'elaborazione dei dati	4
Comprensione dei dati	5
Descrizione dei dati.....	5
Analisi dei dati	6
Qualità dei dati.....	10
Preparazione dati.....	13
Modellizzazione	15
Algoritmi adottati.....	15
Naive bayes	15
Albero decisionale	15
Random forest.....	15
CN2:	16
Algoritmo maggioritario:.....	16
Partizionamento dataset & Indici di valutazione	16
Tuning dei parametri	17
Random forest.....	17
Albero decisionale	17
CN2	18
Risultati.....	18
Valutazione dei risultati	19
Strategia alternativa	19
Comprensione dei dati 2	19
Preparazione dati 2.....	21
Modellizzazione 2	22
Algoritmi aggiuntivi adottati	22
Regressione logistica	22
Support vector machine.....	25
k-nearest neighbour.....	25
Partizionamento dataset & Indici di valutazione	25
Tuning dei parametri	25
Random forest.....	26
Albero decisionale	26
CN2	26
Regressione logistica	27
Support vector machine.....	27
K-nearest neighbour.....	28
Risultati.....	28

Valutazione dei risultati 2	28
Diffusione	31
Appendice 1	32
Appendice 2	33
Codice matlab	33
Mutua informazione	33
Funzione per distribuzione di probabilità cumulativa condizionata dall'attributo.....	33

DEFINIZIONE DEL PROBLEMA

OBBIETTIVO

Frequentemente il medico si trova a dover affrontare casistiche per le quali non è perfettamente preparato, perciò fa uso della sua conoscenza e dell'esperienza pregressa per elaborare una diagnosi sulla base delle informazioni a sua disposizione. D'altra parte, la diagnosi di una data malattia a fronte degli stessi sintomi è operatore dipendente e, di conseguenza, non vi è la certezza che tutti i cardiologi diagnostichino correttamente una malattia cardiaca, se presente. Per cui, un'analisi automatica dei dati che possa fornire un risultato con un grado di affidabilità relativamente alto, potrebbe assistere efficacemente ed efficientemente il medico nelle sue diagnosi. Questo non vuol dire che il medico dovrà sottostare alle indicazioni fornite dal sistema automatico, bensì potrà privilegiare della sua presenza laddove dovesse riscontrare dei dubbi.

Nel nostro paese le malattie cardiovascolari rappresentano la principale causa di morte, essendo responsabili del 44% di tutti i decessi. Inoltre, sul fronte economico, ben il 23,5% della spesa farmaceutica italiana è destinata a farmaci per il sistema cardiovascolare (dati forniti da "epicentro-portale dell'epidemiologia per la sanità pubblica"). Pertanto, migliorando il sistema diagnostico di questa malattia si avrebbero notevoli vantaggi in termini e di sopravvivenza globale e di spesa pubblica.

GOAL DELL'ELABORAZIONE DEI DATI

Il file, contenente 303 esempi di pazienti, è stato fornito dalla Cleveland Clinic Foundation grazie alla collaborazione del dottor Robert Detrano.

Il sistema automatico si propone di elaborare tutte o alcune delle informazioni del singolo paziente e di restituire una diagnosi qualitativa della probabilità che il paziente abbia una malattia cardiaca. Tale probabilità è espressa sotto forma di un numero naturale ordinale compreso tra 0 e 4, dove lo 0 indica che il paziente non è malato e il 4 indica che vi è una forte probabilità che lo sia. I risultati verranno ritenuti significativi se l'accuratezza, in termini assoluti del classificatore, supererà l'80%, ovvero se riuscirà a predire correttamente lo stato di salute del paziente nell'80% dei casi.

Al fine di produrre il miglior risultato possibile, si calcoleranno gli score dei maggiori classificatori disponibili ad oggi nel campo del machine learning, tra i quali naive bayes, regressione logistica, random forest. Quindi, si effettueranno delle apposite procedure di analisi statistica che indicheranno quale degli algoritmi utilizzati possa essere quello più adatto alla situazione da noi considerata.

COMPRENSIONE DEI DATI

DESCRIZIONE DEI DATI

Il dataset fornito contiene 303 record e 13 attributi più la classe. Di seguito è riportata la tabella in cui viene specificato il nome, il formato e una breve descrizione di ogni singolo attributo presente nel dataset.

ATTRIBUTI		
NOME	TIPO	DESCRIZIONE
Age	Continuo	Età in anni
Sex	Binario	0 = donna 1 = uomo
Cp	Discreto	Dolore al petto: 1=angina tipico, 2=angina atipico, 3=dolore non dovuto ad angina, 4=asintomatico <i>L'angina pectoris è il dolore toracico che si verifica quando c'è un limitato trasporto di ossigeno al muscolo cardiaco.</i>
Trestbps	Continuo	Pressione arteriosa sistolica [mmHg]
Chol	Continuo	Livello di colesterolo [mg/dl] <i>È il "serum cholesterol" ovvero il colesterolo totale. È una combinazione del colesterolo HDL ed LDL.</i>
Fbs	Binario	Livello di zuccheri nel sangue a digiuno>120mg/dl: 1=vero, 0=falso
Restecg	Discreto	Ecg a riposo: 0=normale, 1=anormalità dell'onda ST-T, 2=probabile ipertrofia ventricolare sinistra
Thalach	Continuo	Massimo rate cardiaco raggiunto
Exang	Binario	Lo sforzo induce angina: 1=sì, 0=no
Oldpeak	Continuo	Depressione onda ST dovuta all'esercizio da fermo
Slope	Discreto	Pendenza del picco onda ST durante l'esercizio: 1=pendenza positiva, 2=piatta, 3=pendenza negativa
Ca	Discreto	Numero di vasi evidenziati dalla fluoroscopia: 0, 1, 2, 3.
Thal	Discreto	Esito esame di scintigrafia miocardica con tracciante Tallio 201 cloruro: 3 = normale, 6 = fixed defect → anomalia presente a riposo e sotto sforzo 7 = reversable defect → anomalia presente solo sotto sforzo

CLASSE		
NOME	TIPO	ATTRIBUTO
Num	Discreto	Diagnosi malattia cardiaca: 0 = No 1=bassa probabilità 2≥1 3≥2 4=alta probabilità

ANALISI DEI DATI

Prima di effettuare modifiche o migliorie all'interno del dataset, è opportuno valutare come sono distribuiti gli attributi e le relazioni fra essi, attraverso analisi puramente statistiche. In prima battuta si valutano individualmente gli attributi, fornendone una misura di tendenza centrale e, laddove possibile, una misura di variabilità.

ATTRIBUTI CONTINUI		
NOME	MEDIA(troncata alla seconda cifra decimale)	DEVIAZIONE STANDARD CAMPIONARIA
Age	54.43	9.03
Trestbps	131.689	17.599
Chol	246.693	51.776
Thalach	149.607	22.875
Oldpeak	1.039	1.161

ATTRIBUTI DISCRETI				
NOME	MODA	VALORI ASSUNTI	FREQ. RELATIVA	FREQ. CUMULATA
Sex	1	0	0.32	0.32
		1	0.68	1
Cp	4	1	0.07	0.07
		2	0.17	0.24
		3	0.29	0.53
		4	0.47	1
Fbs	0	0	0.85	0.85
		1	0.15	1
Restecg	0	0	0.50	0.50
		1	0.01	0.51
		2	0.49	1
Exang	0	0	0.68	0.68
		1	0.32	1
Slope	1	1	0.47	0.47
		2	0.46	0.93
		3	0.07	1
Thal	3	3	0.55	0.55
		6	0.06	0.61
		7	0.39	1
Ca	0	0	0.59	0.59
		1	0.22	0.81
		2	0.13	0.94
		3	0.06	1

Analizzando qualitativamente le distribuzioni di ogni singola variabile e le tabelle relative, sono emersi i seguenti risultati: l'attributo fbs è relativamente stabile con una bassa varianza, pertanto ci si potrebbe aspettare che l'influenza di tale attributo risulti meno significativa delle altre sulla predizione della classe.

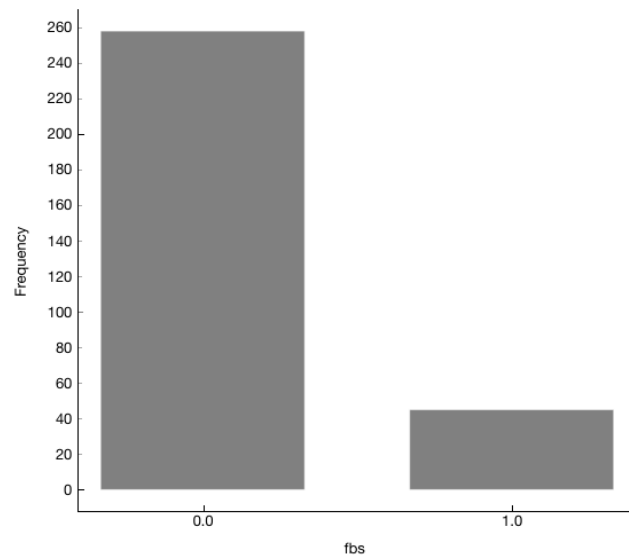


FIGURA 1: DISTRIBUZIONE FREQUENZE FBS

Inoltre, osservando gli attributi restecg, slope e thal, risulta che alcuni valori sono nettamente più frequenti all'interno del dataset considerato.

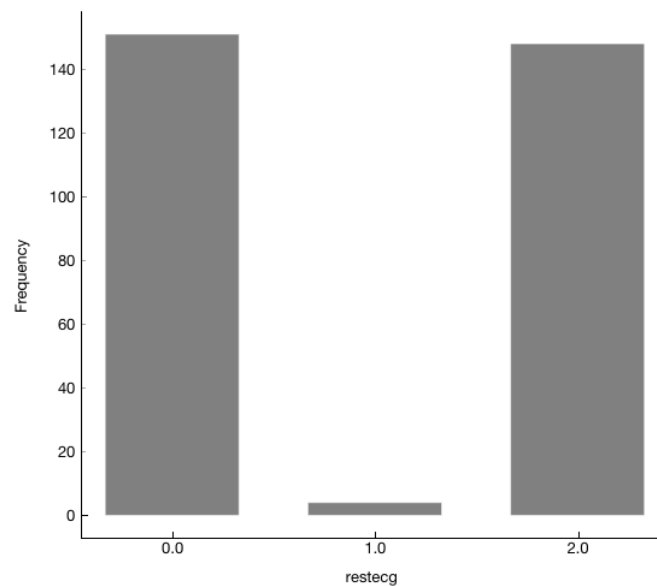


FIGURA 2: DISTRIBUZIONE FREQUENZE RESTECG

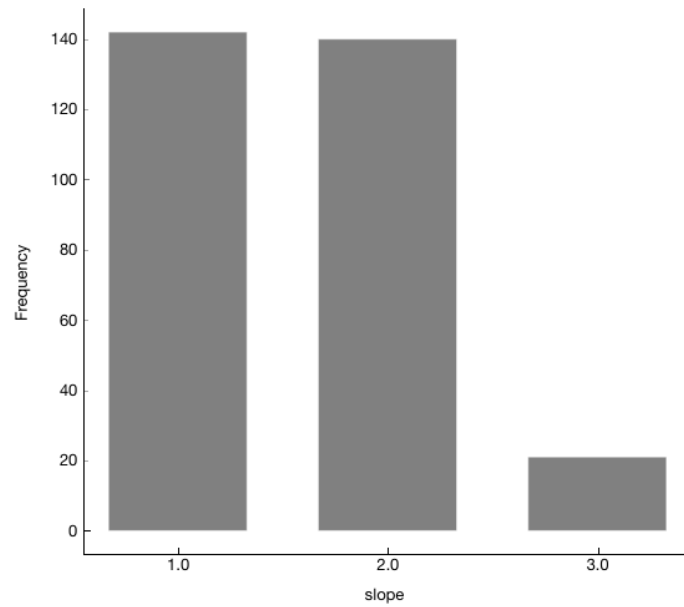


FIGURA 3: DISTRIBUZIONE FREQUENZE SLOPE

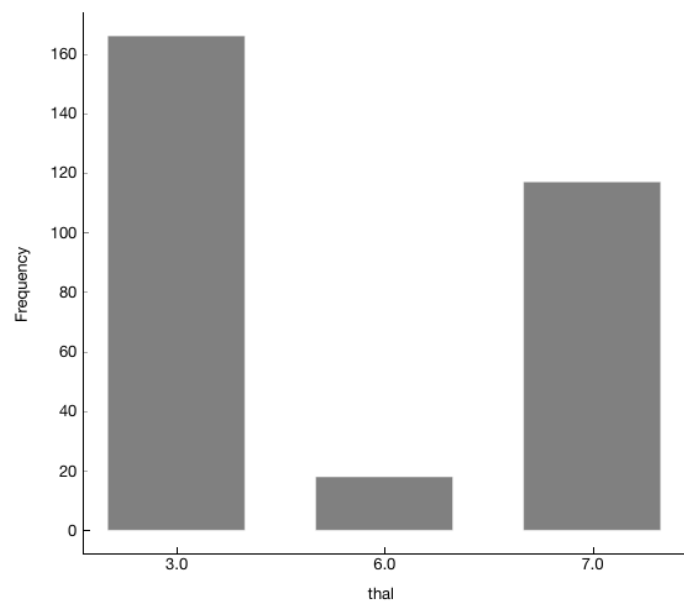


FIGURA 4: DISTRIBUZIONE FREQUENZE THAL

Per quanto riguarda le altre variabili, invece, non sono state riscontrate sostanziali differenze sulle frequenze dei valori.

Dopodiché, osservando i boxplot delle variabili continue, non è stata rilevata la presenza di outlier.

Non meno importante è la valutazione delle frequenze della classe, riportata di seguito:

CLASSE	TIPO	VALORI	FREQ. RELATIVA
Num	Discreto	0	0.54
		1	0.18
		2	0.12
		3	0.12
		4	0.04

È interessante notare quanto sbilanciata sia la distribuzione delle frequenze, la quale potrebbe compromettere le prestazioni dei classificatori.

QUALITÀ DEI DATI

Un'ulteriore indicatore della qualità dei dati viene fornito dalla percentuale di dati mancanti, qualora ve ne siano. Essa si aggira nel nostro dataset sullo 0,2% e, in particolare, solo due delle tredici variabili sono interessate da questo fenomeno, *ca* e *thal*. La prima corrisponde al numero di vasi evidenziati dalla fluoroscopia, mentre la seconda corrisponde all'esito di un esame scintigrafico. Essendo due tecniche che utilizzano rispettivamente raggi X e radiofarmaci, presentano un potenziale rischio per la salute del paziente. Pertanto, si potrebbe immaginare che un cardiologo, confidente della buona salute del paziente, decida di non sottoporre quest'ultimo a questa tipologia di esami, perché non ritenuti necessari.

In realtà, considerata la scarsissima frequenza di dati mancanti, è difficile poter affermare che ci sia una qualche sistematicità che influenza la loro assenza. Pertanto, si ipotizza che i dati siano missing at random, ovvero mancano in modo casuale.

Dopo aver valutato singolarmente gli attributi, è opportuno svolgere un'analisi atta a verificare se vi è correlazione fra essi. La natura di alcuni dati suggerisce che alcuni di essi possano essere estremamente correlati. Si veda ad esempio il *thal* e il *cp* che sono, rispettivamente, il risultato del test scintigrafico e il dolore al petto. Infatti, un dolore al petto atipico è, nel 68% dei casi, legato ad un esito positivo del test scintigrafico. Inoltre, la diagnosi medica condotta sulla base delle variabili legate all'ecg sotto sforzo ha, tipicamente, una sensibilità e specificità paragonabili a quella ottenuta dal test scintigrafico, pertanto si immagina che esse siano correlate. A tal proposito si impiega l'indice di mutua informazione, che richiede dati discreti. Per cui i dati continui all'interno del dataset vengono discretizzati in modo non supervisionato.

Si è scelto di partizionare il range dei valori assunti dagli attributi continui in 4 intervalli, suddivisi in egual frequenza. Le soglie ottenute dall'operazione di discretizzazione sono le seguenti:

NOME	SOGLIA1	SOGLIA2	SOGLIA3
Age	47.50	55.50	60.50
Trestbps	119.00	129.50	139.00
Chol	211.50	241.50	275.50
Thalach	133.50	152.50	165.50
Oldpeak	0.05	0.95	1.85
Ca	0.50	1.50	2.50

In seguito si è effettuato un inputing con cui si è sostituito al dato mancante una misura di tendenza centrale dell'attributo in questione.

Si è, quindi, in grado di calcolare la mutua informazione per ogni coppia di attributi. I risultati sono riportati di seguito:

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal
Age	1.999	0.0146	0.0399	0.0799	0.0326	0.0215	0.0237	0.1258	0.0087	0.0642	0.0276	0.1364	0.0674
Sex		0.9045	0.0166	0.0126	0.0178	0.0017	0.0079	0.0102	0.0160	0.0132	0.0018	0.0129	0.1189
Cp			1.7370	0.0390	0.0189	0.0095	0.0251	0.1232	0.1670	0.1325	0.0643	0.0862	0.1013
Trestbps				1.9763	0.0281	0.0134	0.0180	0.0229	0.0046	0.0277	0.0129	0	0.0443
Chol					1.9999	0.0021	0.0432	0.0225	0.0083	0.0190	0.0033	0.0301	0.0161
Fbs						0.6061	0.0066	0.0065	0.0004	0.0059	0.0067	0.0105	0.0110
Restecg							1.0881	0.0302	0.0065	0.0395	0.0271	0.194	0.0057
Thalach								1.9999	0.1192	0.1669	0.1753	0.0813	0.0989
Exang									0.9116	0.0638	0.0618	0.0273	0.0794
Oldpeak										1.9772	0.3096	0.0754	0.1192
Slope											1.2940	0.0273	0.0878
Ca												1.6485	0.0595
Thal													1.2955

Si riscontra che nessuna delle mutue informazioni risulta essere significativa, in quanto nessuna supera il valore limite di 1. A conferma di ciò, un'analisi esplorativa degli scatter plot ha prodotto gli stessi risultati. Di conseguenza, possiamo ritenere che i dati non sono correlati.

Inoltre, al fine di confermare la casualità dei dati mancanti, si è sostituito al valore mancante un nuovo valore. Si è, quindi, ricalcolata la mutua informazione. I dati potenzialmente interessati da questo fenomeno sono il Ca e il Thal.

Si riportano i soli valori della mutua informazione calcolati per le due variabili e nei due modi (inputing o sostituzione), così da poter apprezzare eventuali sistematicità nella mancanza del dato.

	Ca (inputing)	Ca (sostituzione)	Thal (inputing)	Thal (sostituzione)
Age	0.1364	0.1364	0.0674	0.0674
Sex	0.0129	0.0216	0.1189	0.1189
Cp	0.0862	0.0892	0.1013	0.1013
Trestbps	0	0.0313	0.0443	0.0443
Chol	0.0301	0.0358	0.0161	0.0161
Fbs	0.0105	0.0164	0.0110	0.0110
Restecg	0.0194	0.0212	0.0057	0.0057
Thalach	0.0813	0.0926	0.0989	0.0989
Exang	0.0299	0.0299	0.0794	0.0794
Oldpeak	0.0754	0.0849	0.1192	0.1192
Slope	0.0273	0.0292	0.0878	0.0878
Ca	1.6485	1.6485	0.0595	0.0595
Thal	0.0273	0	1.2955	1.2955

Si nota che non vi è differenza significativa tra le mutue informazioni, calcolate prima e dopo la sostituzione dei dati mancanti. Viene confermata, quindi, l'ipotesi di dati missing at random.

PREPARAZIONE DATI

Dopo un'attenta analisi non supervisionata, è stato riscontrato che alcuni attributi risultano essere meno variabili di altri. Nonostante questo, non vi è abbastanza certezza che questi attributi possano essere eliminati dal dataset. Perciò si valutano, ed eventualmente vengono eliminati, a seguito di un'analisi supervisionata.

Ora, usufruendo delle distribuzioni osservate prima (*Figura 2*), si valuta se può essere utile effettuare accorpamenti tra le variabili considerate. Si nota, infatti, che il dato più critico da questo punto di vista è il *Restecg*. Esso è un dato nominale che può assumere i valori 0, 1 e 2. Il valore 1, che indica un'anormalità dell'onda ST, ha una frequenza notevolmente inferiore rispetto agli altri due e l'accorpamento con il valore 2, il quale indica a sua volta un'anormalità maggiore, è più sensato in quanto rimarcherebbe la differenza tra ecg normale e anormale. I grafici, prima e dopo l'accorpamento, sono i seguenti:

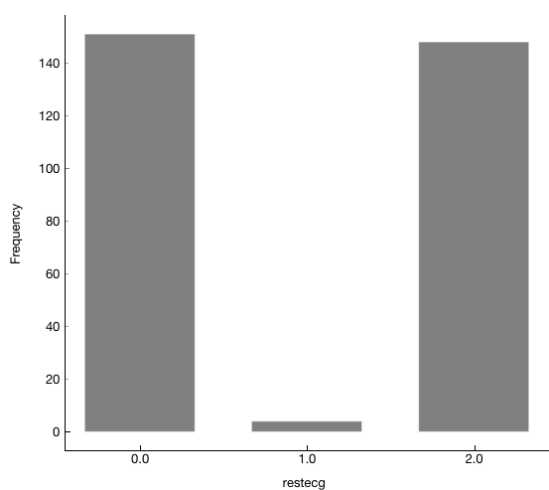


FIGURA 5: DISTRIBUZIONE FREQUENZE RESTECG, SENZA ACCORPAMENTO

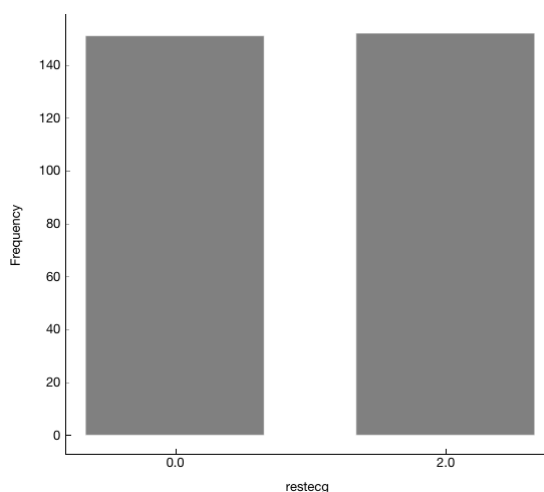


FIGURA 6: DISTRIBUZIONE FREQUENZE RESTECG, CON ACCORPAMENTO














La stessa procedura potrebbe essere adottata anche per gli attributi Slope e Thal. Tuttavia si è preferito evitare di effettuare accorpamenti su tali dati in quanto il valore a minor frequenza risulta essere pari a 18.

A questo punto, è opportuno rivalutare se vi è mutua informazione tra gli attributi. L'accorpamento di *Restecg* potrebbe, infatti, comportare cambiamenti sui valori di MI calcolati precedentemente.

	Restecg (non accorpato)	Restecg(accorpato)
age	0.0237	0.0224
Sex	0.0079	0.0048
Cp	0.0251	0.0248
Trestbps	0.0180	0.0171
Chol	0.0432	0.0343
Fbs	0.0066	0.0077
Restecg	1.0881	1.0014
Thalach	0.0302	0.0216
Exang	0.0065	0.0105
Oldpeak	0.0395	0.0290
Slope	0.0271	0.0212
Ca	0.194	0.0170
Thal	0.0057	0.0071

Dai risultati ottenuti si riscontra che l'accorpamento dei valori dell'attributo *restecg* non ha prodotto cambiamenti significativi in termini di mutua informazione.

Ora, adottando un approccio supervisionato sul 30% dei dati, attraverso una strategia di ranking, si assegna ad ogni attributo un indice di qualità. Come indice si sceglie il gain ratio, data la presenza di attributi misti (discreti e continui). Quindi li ordiniamo dal migliore al peggiore e otteniamo la seguente tabella dei ranks.

	#	Gain Ratio ▼
 ca	4	<u>0.230</u>
 slope	3	<u>0.204</u>
 cp	4	<u>0.185</u>
 thal	3	<u>0.177</u>
 exang	2	<u>0.161</u>
 thalach	C	<u>0.149</u>
 oldpeak	C	<u>0.146</u>
 fbs	2	<u>0.112</u>
 restecg	2	<u>0.101</u>
 age	C	<u>0.100</u>
 sex	2	<u>0.059</u>
 chol	C	<u>0.051</u>
 trestbps	C	<u>0.031</u>

I risultati ottenuti, pur essendo poco significativi in alcuni casi, non consentono di eliminare uno o più attributi. Infatti, un information gain del 3%, come avviene nel *trestbps*, potrebbe costituire un aiuto, seppur esiguo, al classificatore che adotta quell'attributo.

MODELLIZZAZIONE

In questa sezione vengono descritti sommariamente gli algoritmi ritenuti efficaci per il dataset, specificandone la logica di funzionamento che sta alla base e le assunzioni fatte sui dati.

ALGORITMI ADOTTATI

NAIVE BAYES

Si tratta di un metodo generativo che fa uso della statistica bayesiana. È molto apprezzato e utilizzato in ambito clinico in quanto permette di quantificare il contributo di ogni singola variabile sul valore della classe. Esso fa la forte assunzione che gli attributi siano indipendenti e classifica facendo uso di una stima MAP (massimo a posteriori) della classe dati i valori degli attributi. Inoltre, esso presuppone che i dati siano discreti (o siano stati discretizzati) e non fa assunzioni distribuzionali sulle variabili. Lo svantaggio principale di tale algoritmo è che soffre della cosiddetta *curse of dimensionality*. Infatti, un numero di attributi troppo alto porterebbe a risultati disastrosi della stima della probabilità a posteriori.

ALBERO DECISIONALE

È un algoritmo induttivo caratterizzato da un basso bias e un'alta varianza, che tipicamente lo porta a performance non paragonabili a quelle degli algoritmi concorrenti. Nonostante questo, è molto utilizzato. Infatti, l'algoritmo TDIDT di cui fa uso, accoppiato all'adozione di regole if-then, lo rende un ottimo descrittore del dataset. Ha, inoltre, la capacità di ordinare gli attributi per importanza misurata con l'information gain, e di scartarli, eventualmente, adottando delle tecniche di prepruning o di pruning. Nel caso più generale, adotta un decisore probabilistico che classifica con una stima MAP della classe dati gli attributi. Infine, ha il vantaggio di essere utilizzabile sia con attributi discreti che continui ed esiste una variante dell'albero, chiamata regression tree, che permette di lavorare anche con classi continue.

RANDOM FOREST

È un metodo discendente dagli alberi decisionali, da cui eredita l'impostazione e il basso bias. Si basa sull'idea di mediare le soluzioni ottenute da N alberi decisionali al fine di minimizzare la varianza del classificatore. Così come nel singolo albero, può essere utilizzato con attributi di qualsiasi tipo. L'algoritmo classifica andando a massimizzare la probabilità a posteriori della classe ottenuta dagli alberi decisionali, e andando a mediare la soluzione degli alberi di regressione.

L'algoritmo fa tali assunzioni:

- utilizza $p < m$ attributi per effettuare una diramazione

- costruisce alberi binari
- non fa pruning
- può effettuare prepruning
- i dati utilizzati per l'apprendimento del singolo albero sono campioni di bootstrap del dataset originale, aventi la stessa numerosità di tale dataset
- conosce i parametri del modello: numero di alberi costruiti, numero di attributi utilizzabili durante la diramazione, eventuali parametri di prepruning

Inoltre, questo algoritmo ha la capacità di assegnare un ranking alle variabili, selezionando così quelle più importanti al momento del runtime. Tali caratteristiche rendono l'algoritmo il migliore sul mercato quando la numerosità degli attributi è molto alta e non vi è possibilità di scartare a priori alcune delle variabili considerate.

CN2:

Algoritmo di facile implementazione e comprensione, basato sull'apprendimento di regole if-then valutate con confidenza, supporto e copertura. Fa parte dei cosiddetti algoritmi di covering da cui eredita la sua greedness, per cui fa spesso uso di una beam search per supplire a questa mancanza. Sebbene le sue performance non siano eccellenti nella maggior parte dei casi, è consigliato utilizzarlo in quanto è un ottimo descrittore del dataset.

ALGORITMO MAGGIORITARIO:

Algoritmo più semplice adottabile, fa votare tutti gli esempi e sceglie la classe che ha una maggiore frequenza all'interno del training set. E' spesso associato ad accuratezze molto basse, nonostante raggiunga valori estremamente alti quando la distribuzione della classe è fortemente sbilanciata. Valori alti risultano comunque poco affidabili nel caso in cui lo scopo sia quello di predire la classe a frequenza minore. È comunque utile confrontarlo con gli altri algoritmi, usandolo come riferimento.

PARTIZIONAMENTO DATASET & INDICI DI VALUTAZIONE

La procedura di test è effettuata sul 70% di dati non utilizzati per la feature selection. Consiste nel valutare alcuni indici di qualità dei classificatori ottenuti dopo una procedura di 3-fold cross validation a fold stratificati per classe. La scelta di soli 3 fold è dovuta alla scarsità della classe num 4, in quanto la stratificazione per classe su un numero di fold maggiore implicherebbe l'assenza di tale valore della classe nei test set. Avendo 5 possibili valori della classe, si è scelto di adottare l'accuratezza di generalizzazione. Ovviamente, la procedura di cross-validazione comporta la suddivisione del dataset in un training-set e in un test-set, che variano ad ogni iterazione dell'algoritmo di validazione, così da poter beneficiare di 3 test-set indipendenti.

Dal momento che alcuni algoritmi richiedono che i dati siano preventivamente discretizzati e normalizzati, questa operazione viene effettuata su tutti i dati continui all'interno del dataset, ma solo ed esclusivamente se l'algoritmo lo richiede. Per cui, si è deciso di non discretizzare tutti i dati indistintamente, altrimenti si rischierebbe di perdere dell'informazione importante. In particolare, la

discretizzazione è stata effettuata partizionando i dati in 10 intervalli di egual frequenza. I dati sono stati, inoltre, normalizzati rispetto alla deviazione standard.

TUNING DEI PARAMETRI

Alcuni degli algoritmi adottati necessita del settaggio dei parametri. Naturalmente questo deve essere effettuato all'interno di un validation-set indipendente, dove vengono testati dei valori plausibili del parametro considerato, allo scopo di incrementare la qualità della classificazione in termini di accuratezza di generalizzazione. Purtroppo le dimensioni del dataset non consentono di suddividerlo ulteriormente, per cui si adotta come validation-set lo stesso dataset utilizzato per il ranking supervisionato. Saranno eventualmente proposte in questo report delle tabelle risultanti da una grid-search dei parametri, così da facilitare futuri riutilizzi del dataset heart_disease_cleveland.

RANDOM FOREST

È necessario stabilire il numero (p) di attributi con il quale effettuare ogni split del singolo albero. Per convenzione, si sceglie p uguale alla radice del numero di attributi. Per cui, avendo 13 attributi scegliamo 4, ovvero l'intero più vicino alla radice calcolata. Inoltre, come tecnica di prepruning, la crescita dell'albero viene fermata quando la numerosità degli esempi è inferiore o uguale a 5. Questo è frutto di una grid search sulle accuratezze al variare della numerosità da 2 a 11. Ovviamente si è scelto il valore che apporta la migliore accuratezza di generalizzazione.

Dimensione limite del subset	2	3	4	<u>5</u>	6	7	≥ 8
CA	0.604	0.604	0.604	<u>0.615</u>	0.593	0.571	≤ 0.571

Infine, sebbene a un maggior numero di alberi corrisponda una varianza inferiore, si è scelto di far crescere 40 alberi, in quanto un numero troppo alto comporterebbe la creazione di bootstrap sample estremamente dipendenti fra loro.

ALBERO DECISIONALE

Anche in questo caso si adotta una tecnica di prepruning influenzata dal valore dell'accuratezza.

Dimensione limite del subset	≤ 9	10	<u>11</u>	≥ 12
CA	≤ 0.484	0.505	<u>0.516</u>	≤ 0.516

A seguito di una grid search si è scelto 11 come dimensione minima dei dati.

CN2

Questo algoritmo necessita del settaggio della dimensione del fascio di ricerca delle regole decisionali e della copertura minima come scelta implicita della lunghezza delle regole. Si è, quindi, effettuata una doppia grid-search, la prima atta a trovare la dimensione subottima del fascio, la seconda per trovare il valore di copertura subottimo della regola. In orange è possibile, inoltre, scegliere una lunghezza massima delle regole. Avendo svolto questa mansione sul valore minimo della copertura, si è ignorata la scelta di questo ulteriore parametro, settato a 10 di default.

Per quanto riguarda le dimensioni del fascio, sembrerebbe che esso non pregiudichi i risultati in termini di accuratezza, pertanto si è deciso di fissarlo a 5, ovvero al valore di default.

Per quanto riguarda invece la copertura minima delle regole, quando questa è uguale a 0,11 si nota un netto miglioramento dell'accuratezza.

Copertura	<0.05	0.05	<u>0.11</u>	0.16	0.22	>0.22
CA	<0.429	0.429	<u>0.527</u>	0.505	0.462	≤0.462

Osservazione: Ad ogni run di orange si ha un rimescolamento dei dati di training e di test all'interno della cross validazione. Pertanto, seppur di poco, le accuratezze ottenute a seguito di un tuning dei parametri considerati, potrebbe variare. Si è riscontrato che questo avviene frequentemente, in particolare per l'algoritmo random forest, e che non sempre il valore 5 può essere ritenuto il migliore. È doveroso, dunque, far presente che i valori settati sono il risultato di un tuning subottimo.

RISULTATI

I risultati ottenuti al termine della procedura di cross validazione sono i seguenti:

Algoritmo	Maggioritario	Albero decisionale	Random forest	Naive bayes	CN2
CA-media	0.545	0.557	0.587	0.534	0.522

VALUTAZIONE DEI RISULTATI

Si osserva quanto siano bassi i valori di accuratezza ottenuti e che il goal (accuratezza \geq 80%) non è neanche lontanamente rispettato. La natura della classe ci suggerisce, però, una sostanziale modifica al dataset, la quale potrebbe portare notevoli benefici in termini di qualità di classificazione. La classe, infatti, oltre ad indicare la probabilità qualitativa di avere la malattia, indica, sempre qualitativamente, la percentuale di stenosi dei maggiori vasi cardiaci. Se la classe è 0 vi è una percentuale di stenosi minore del 50%, se è maggiore o uguale a 1 allora la percentuale di stenosi è maggiore del 50% e cresce con l'aumentare del valore intero della classe. Inoltre, è evidente dalla letteratura che un'occlusione maggiore del 50% comporti seri rischi per la salute del paziente, il quale venendo classificato come affetto da malattie cardiache, deve sottoporsi ad un intervento chirurgico per ripristinare l'afflusso di sangue nel vaso interessato.

STRATEGIA ALTERNATIVA

La binarizzazione della classe atta a distinguere i casi in cui la stenosi è minore del 50% e i casi opposti potrebbe essere la strategia più vantaggiosa.

CLASSE		
NOME	TIPO	DESCRIZIONE
Num	Discreto	Diagnosi: 0: stenosi dei maggiori vasi cardiaci <50%, 1: stenosi dei maggiori vasi cardiaci>50%

Questa operazione comporta il rifacimento di alcuni punti già svolti, quali la feature selection supervisionata, il tuning dei parametri e l'introduzione di nuovi algoritmi.

D'ora in avanti si considereranno le stesse porzioni dei dati sopra descritte per training-set, test-set e validation-set.

COMPRENSIONE DEI DATI 2

Si è effettuata una valutazione delle frequenze della classe, le quali potrebbero indicare quali indici di qualità calcolare nella fase di modellizzazione e valutazione.

CLASSE	TIPO	VALORI	FREQ. RELATIVA
Num	Discreto	0	0.54
		1	0.46

Tali frequenze suggeriscono che la scelta dei pazienti all'interno del dataset non sia stata effettuata in maniera casuale su un campione di soggetti appartenenti alla popolazione. Sarebbe impensabile, infatti, che il 46% della popolazione sia affetta da malattia cardiaca. A conferma di questa considerazione, si ricordano i risultati ottenuti in uno studio di popolazione americano di Nkomo et al. (2006) in cui si valuta l'impatto sulla prevalenza di tale malattia al variare dell'età dei pazienti, senza considerarne il sesso. In particolare, i soggetti di età compresa tra i 45 e 54 anni riscontrano una prevalenza complessiva dello 0.4%, i soggetti di età compresa tra i 55 e i 64 anni hanno una prevalenza complessiva dell'8.5%. Sono stati presi in considerazione queste fasce di età in quanto la media delle età dei pazienti contenuti nel data set considerato è di 54,44 anni.

Tale considerazione consiglia di effettuare una correzione in fase di modellizzazione, sostituendo la probabilità a priori degli algoritmi bayesiani con la prevalenza reale di popolazione per la relativa area geografica.

A questo punto, l'analisi visiva dei cluster effettuata con l'MDS ha permesso di trarre la seguente conclusione. Sebbene è evidente che vi sia un solo cluster inscindibile, colorando la classe è chiaro che le classi sono separabili da un iperpiano, seppur con qualche errore. Pertanto ci si aspetta che algoritmi che ipotizzano la separabilità dei dati rispondano bene in fase di apprendimento.

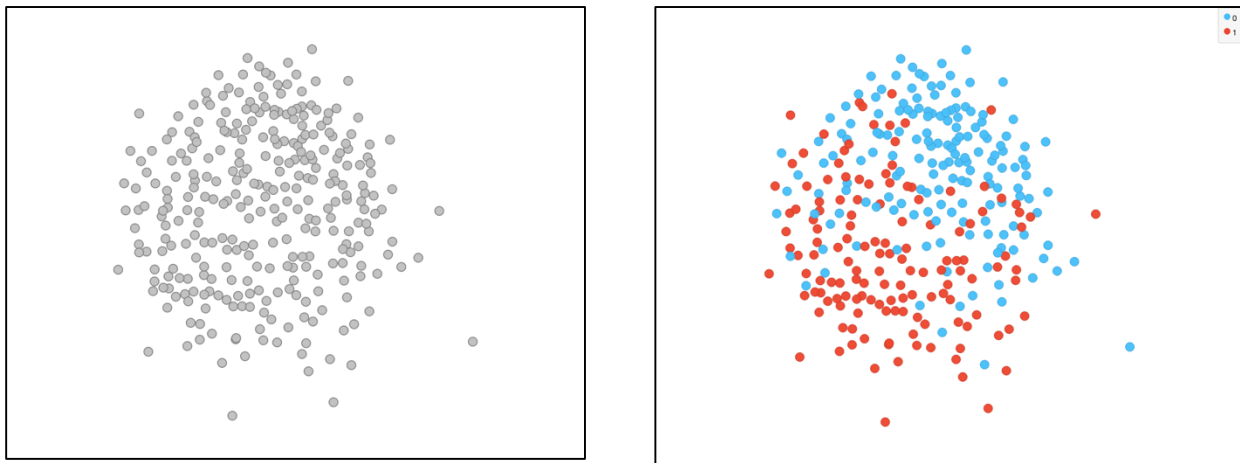


FIGURA 8: A SINISTRA CLUSTER SENZA CLASSE. A DESTRA CLUSTER CON CLASSE

A conferma della separazione indicata dall'MDS, l'analisi esplorativa svolta con un algoritmo 2-means mostra relazioni importanti tra i cluster trovati e la classe reale, nonostante ci siano comunque degli errori. Questo indica che l'algoritmo riesce comunque a trovare delle relazioni naturali tra gli esempi che possano indicare due cluster differenti, che coincidono con le classi reali.

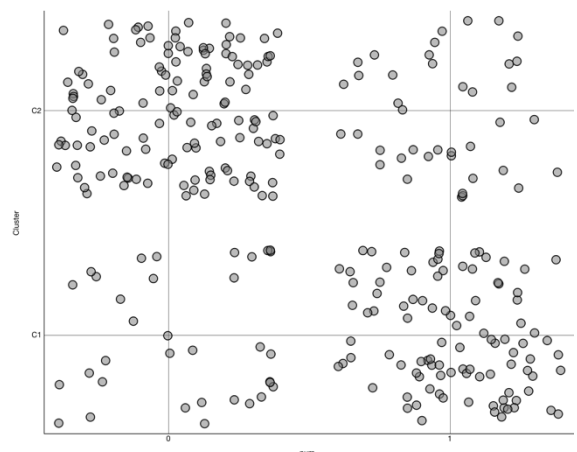















FIGURA 7: SCATTER PLOT CLUSTER-CLASSE

PREPARAZIONE DATI 2

La feature selection supervisionata ha prodotto i seguenti risultati:

	#	Gain Ratio ▼
 exang	2	0.164
 slope	3	0.133
 thal	3	0.128
 ca	C	0.109
 oldpeak	C	0.102
 cp	4	0.094
 thalach	C	0.058
 restecg	2	0.031
 age	C	0.021
 chol	C	0.019
 sex	2	0.019
 fbs	2	0.014
 trestbps	C	0.013

Si è scelto lo 0,02% come valore di soglia di information gain sotto il quale non considerare le variabili. Pertanto, dato lo scarso punteggio assegnato alla *trestbps*, *fbs*, *sex* e *chol*, si è deciso di non far adottare queste variabili dai classificatori, in quanto non produrrebbero benefici significativi in termini di qualità di classificazione.

Osservazione: Nonostante sia noto dalla letteratura che il sesso influenza la presenza di malattie cardiache, i risultati ottenuti dalla feature selection indicano quanto poco questa variabile condiziona il valore della classe. Questo a conferma del fatto che il dataset è fortemente polarizzato.

Viene mostrato, quindi, il dataset finale su cui saranno effettuate le successive operazioni, quali modellizzazione, valutazione e distribuzione.

ATTRIBUTI		
NOME	TIPO	DESCRIZIONE
Age	Continuo	Età in anni
Cp	Discreto	Dolore al petto: 1=angina tipico, 2=angina atipico, 3=dolore non dovuto ad angina, 4=asintomatico <i>L'angina pectoris è il dolore toracico che si verifica quando c'è un limitato trasporto di ossigeno al muscolo cardiaco.</i>

Restecg	Discreto	Ecg a riposo: 0=normale, 1=anormalità dell'onda ST-T, 2=probabile ipertrofia ventricolare sinistra
Thalach	Continuo	Massimo rate cardiaco raggiunto
Exang	Binario	L'esercizio induce angina: 1=sì, 0=no
Oldpeak	Continuo	Depressione onda ST dovuta all'esercizio da fermo
Slope	Discreto	Pendenza del picco onda ST durante l'esercizio: 1=pendenza positiva, 2=piatta, 3=pendenza negativa
Ca	Discreto	Numero di vasi evidenziati dalla fluoroscopia: 0, 1, 2, 3.
Thal	Discreto	Esito esame di scintigrafia miocardica con tracciante Tallio 201 cloruro: 3 = normale, 6 = fixed defect → anomalia presente a riposo e sotto sforzo 7 = reversable defect → anomalia presente solo sotto sforzo

MODELLIZZAZIONE 2

Sapendo che la classe è binaria, è possibile introdurre altri due algoritmi di classificazione: regressione logistica e support vector machine.

ALGORITMI AGGIUNTIVI ADOTTATI

REGRESSIONE LOGISTICA

È un metodo discriminativo pensato esclusivamente per problemi di classificazione a 2 classi. Esso presuppone che la relazione univariata tra la classe e la variabile sia monotona e che il decision boundary sia lineare. La fase di apprendimento si svolge costruendo un problema di ottimizzazione attorno alla funzione di verosimiglianza, il cui risultato porta ad avere delle stime dei log(odds) per ogni singola variabile.

Al fine di valutare l'ipotesi di monotonia all'interno delle relazioni classe-attributo del nostro dataset, si è condotto uno studio della proporzione dei pazienti affetti o non affetti dalla malattia con un dato fattore di rischio. Poiché le variabili prese in considerazione, tranne il *sex* e il *cp*, possono essere considerate ordinali, ha senso verificare questa condizione su ognuna di esse. Inoltre, anche la variabile *exang*, pur essendo binaria, potrebbe indicarci una particolare relazione con la classe.

In base ai risultati ottenuti, sembrerebbe che non tutte le variabili possano essere considerate un fattore di rischio o protezione per i pazienti.

Pur essendo un dataset fortemente polarizzato, in quanto la scelta dei pazienti è stata svolta a seguito di una stratificazione per presenza della malattia, è comunque apprezzabile una lieve variazione monotona della probabilità di averla in presenza del fattore di rischio *age* (*figura 9*).

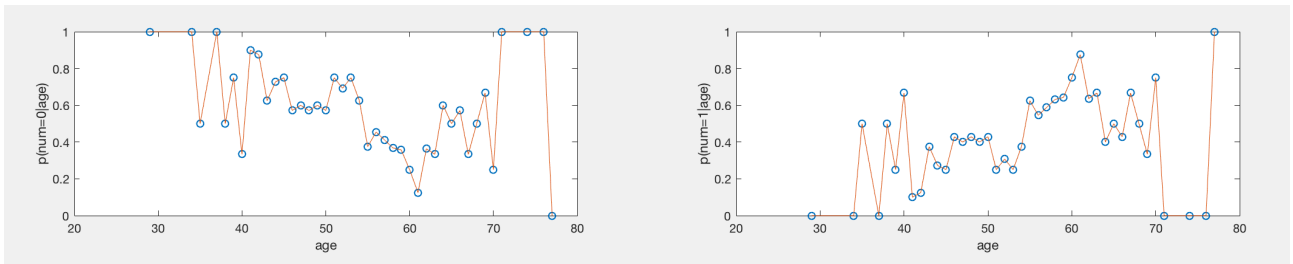


FIGURA 9: RELAZIONE NUM-AGE

Un'anormalità dell'onda ST sembra influenzare la probabilità di avere la malattia, come visibile in figura 10.

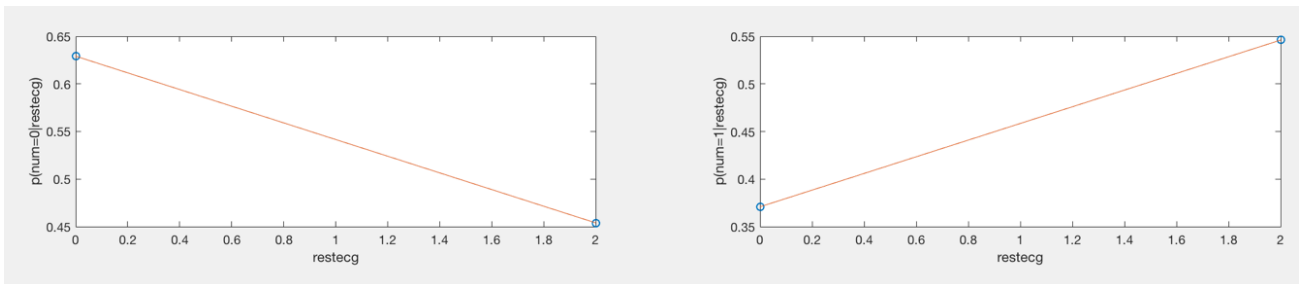


FIGURA 10:RELAZIONE NUM-RESTECG

L'andamento tutto sommato è monotono quando il fattore di rischio è *thalach*, ma fortemente rumoroso (figura 11).

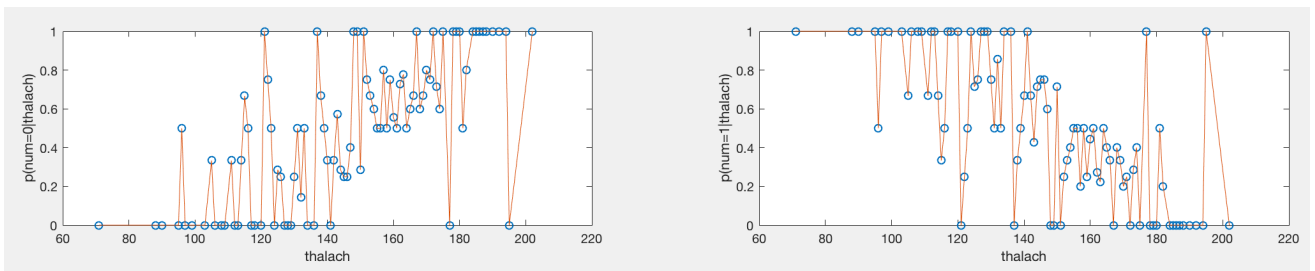


FIGURA 9:RELAZIONE NUM-THALACH

L'andamento è monotono anche per fattore di rischio *exang* (figura 12).

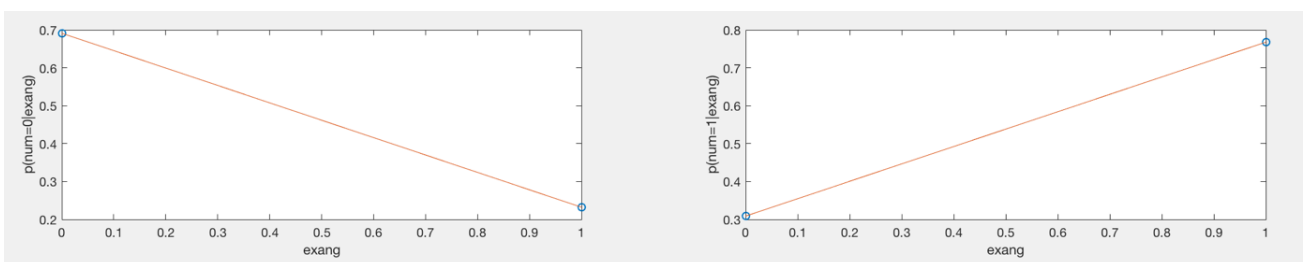


FIGURA 10: RELAZIONE NUM-EXANG

Nonostante la presenza di rumore, è apprezzabile un andamento crescente quando il fattore di rischio *oldpeak* aumenta (figura 13).

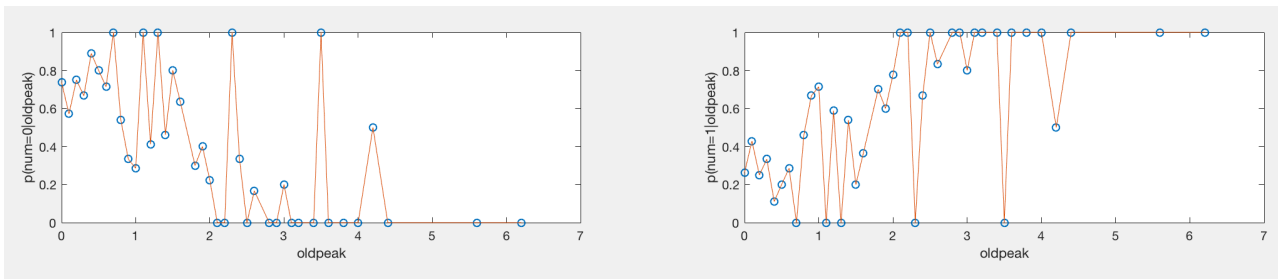


FIGURA 11: RELAZIONE NUM-OLDPEAK

Non vi è evidenza che la relazione num-slope sia monotona (figura 14).

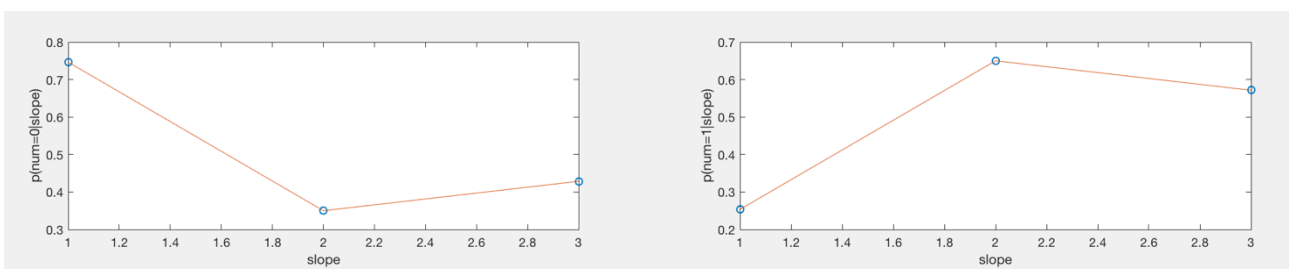


FIGURA 12: RELAZIONE NUM-SLOPE

L'andamento monotono è confermato nel caso in cui si considera *ca* come fattore di rischio (figura 15).

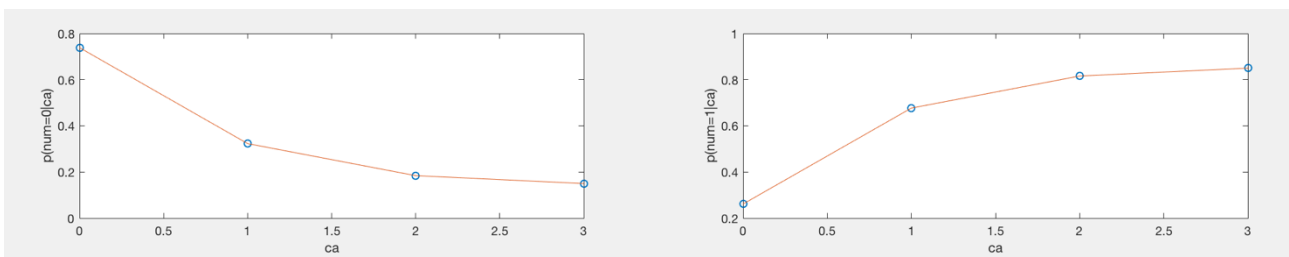


FIGURA 13: RELAZIONE NUM-CA

Anche qui è evidente un andamento monotono (figura 16).

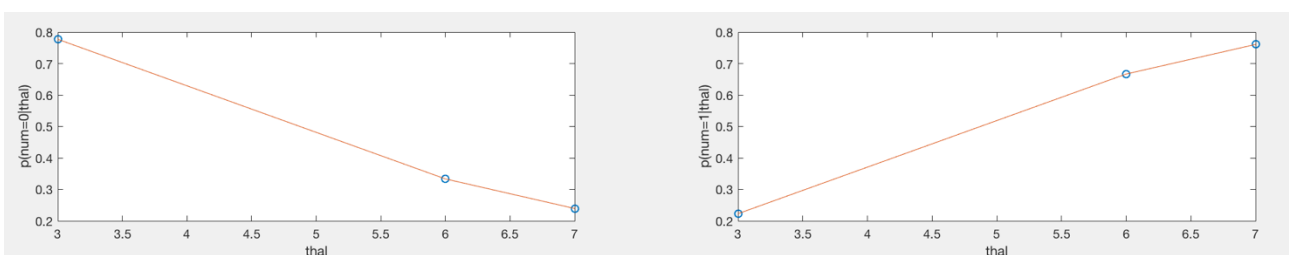


FIGURA 14: RELAZIONE NUM-THAL

In conclusione possiamo affermare di poter utilizzare la regressione logistica in quanto l'assunzione di monotonia è in gran parte verificata.

SUPPORT VECTOR MACHINE

Anche questo è un algoritmo di classificazione pensato per problemi a due classi. Esso presuppone che gli esempi siano linearmente separabili se visti sotto un'opportuna trasformazione dimensionale e risolve la stima dei suoi parametri svolgendo un problema di ottimizzazione vincolata, la cui soluzione permette di esprimere una regola decisionale funzione dell'esempio da classificare.

K-NEAREST NEIGHBOUR

È un approccio molto semplice che sfrutta il concetto di similarità, calcola la distanza degli esempi dal caso che si vuole classificare e fa votare i k più vicini, così da ottenere il valore della classe. Fa, quindi, uso del parametro k , il quale influenza la varianza e il bias del modello, e di una funzione distanza. Questa può essere, a seconda dei casi, utilizzata per calcolare la distanza tra dati continui, discreti o misti. È un approccio interessante in quanto riprende il ragionamento analogico svolto dai medici, che sfrutta l'esperienza pregressa per poter fare delle previsioni. Nell'implementazione in orange è permesso l'utilizzo delle sole distanze relative a dati continui, pertanto si immagina che effettui un calcolo della distanza eterogenea qualora vi siano attributi misti. Sebbene le sue prestazioni in termini di classificazione siano considerate soddisfacenti, è molto lento per grandi quantità di dati, in quanto lavora esclusivamente al momento del runtime.

PARTIZIONAMENTO DATASET & INDICI DI VALUTAZIONE

La procedura di test è effettuata sul 70% di dati rimanenti dalla feature selection. Si sono ricavati degli indici di qualità dei classificatori ottenuti dopo una procedura di 10-fold cross validation. In particolare, sono stati utilizzati indici quali sensibilità, specificità, accuratezza e AUC. Una volta stabilito quale sia l'algoritmo di classificazione migliore, viene costruito un intervallo di confidenza per la sua accuratezza attraverso una statistica t con $k-1$ gdl.

TUNING DEI PARAMETRI

Il settaggio dei parametri deve essere svolto nuovamente dal momento che la classe è stata binarizzata. Vengono testati valori plausibili del parametro considerato allo scopo di aumentare la qualità della classificazione in termini di AUC. Infatti, sebbene l'accuratezza sia un indice di qualità estremamente semplice da calcolare, si preferisce adottare la misura dell'area sotto la curva roc in quanto permette di stabilire l'affidabilità del classificatore al variare della regola decisionale. La necessità di farla variare compare frequentemente in ambito clinico e si sa che essa dipende dal rapporto del costo di sbagliare nei falsi positivi e il costo di sbagliare nei falsi negativi. Non potendo stabilire a priori tale valore, ci si mette in una situazione estremamente generica in cui la soglia decisionale è variabile.

RANDOM FOREST

Come descritto precedentemente, scelgo p pari alla radice della numerosità degli attributi, quindi avendone 9, p risulta pari a 3. Inoltre, stoppata la crescita degli alberi nel caso in cui la numerosità degli esempi sia inferiore o uguale a 8 scegliendo il valore che apporta la migliore AUC. Infine, si è scelto di far crescere 40 alberi.

Dimensione limite del subset	≤ 6	7	<u>8</u>	9	10	≥ 11
AUC	≤ 0.890	0.890	<u>0.911</u>	0.890	0.903	0.895

ALBERO DECISIONALE

Anche qui si adotta una tecnica di prepruning influenzata dal valore dell'AUC. A seguito di una grid search si è scelto 14 come dimensione minima dei dati.

Dimensione limite del subset	<u>≤ 14</u>	≥ 15
AUC	<u>0.750</u>	≤ 0.725

CN2

Lunghezza massima regole=10. Per quanto riguarda le dimensioni del fascio, sembrerebbe che esso non pregiudichi i risultati in termini di AUC. Si è deciso, dunque, di fissare questa dimensione a 5, ovvero al valore di default. Inoltre, come misura di valutazione si è scelta l'entropia, data la presenza di dati continui. Infatti, nonostante sia consigliabile usare una stima di Laplace, l'uso di quest'ultima richiederebbe una discretizzazione preventiva, mentre per quanto riguarda l'entropia, la discretizzazione viene effettuata al fine di massimizzarla. Infine, valutando la copertura minima delle regole, si nota un miglioramento dell'AUC scegliendo 0,16 come copertura minima.

COPERTURA	0.11	<u>0.16</u>	0.22	>0.22
CA	0.674	<u>0.698</u>	0.678	≤ 0.657

REGRESSIONE LOGISTICA

il settaggio dei suoi parametri implementa una particolare forma di wrapping, detta regolarizzazione, in quanto permette di effettuare una selezione naturale degli attributi più importanti per la classificazione. Adotta due diverse tipologie di regolarizzazione: strategia lasso o strategia ridge. Provandole entrambe, al variare del parametro c , si ottengono i seguenti risultati

STRATEGA RIDGE				
c	>0.400	0.180	<u>0.02</u>	<0.02
AUC	<0.905	0.908	<u>0.915</u>	<0.915

STRATEGIA LASSO				
c	>35	7	<u>0.800</u>	<0.800
AUC	<0.814	0.818	<u>0.906</u>	<0.906

Si sceglie, quindi, una strategia ridge con $c=0.02$.

SUPPORT VECTOR MACHINE

Richiede un numero elevato di parametri e di settaggi. Si parte innanzitutto scegliendo la funzione di kernel sulla base dei risultati, tenendo gli altri parametri settati al valore di default.

Si imposta, quindi, la funzione di kernel SIGMOIDE. Dopodiché si realizza una grid search sulla capacità della support vector, il cui risultato subottimo è 1.9.

capacità svm	≤ 1	>1.10 & <1.5	>1.6 & <1.8	<u>1.9</u>	>2
AUC	≤ 0.820	0.827	0.820	<u>0.831</u>	≤ 0.828

A questo punto, si procede in maniera analoga per calcolare i parametri della funzione di kernel. Il parametro g viene tipicamente eguagliato ad $1/k$, dove k è la numerosità degli attributi, ma si hanno prestazioni nettamente superiori ponendolo uguale a 0. Il parametro c settato è anch'esso estratto da una grid search.

c (kernel sigmoide)	<0.90	<u>0.90</u>	>0.90
AUC	≤ 0.838	<u>0.851</u>	≤ 0.840

K-NEAREST NEIGHBOUR

Scelta a priori la distanza euclidea, si effettua una grid search per trovare il valore subottimo di k. Per cui scelgo k=9.

K	≤8	<u>9</u>	≥10
AUC	≤0.895	<u>0.914</u>	≤0.911

RISULTATI

I risultati medi ottenuti al termine della procedura di cross validazione sono i seguenti:

ALGORITMO	ACC. DI GENERALIZZAZIONE	SENS.	SPEC.	AUC
Constant	0.542	0	1	0.5
Albero decisionale	0.774	0.723	0.818	0.862
CN2 rule inductor	0.784	0.734	0.826	0.830
Random forest	0.835	0.774	0.806	0.886
Naive Bayes	0.826	0.783	0.861	0.897
Regressione logistica	0.765	0.714	0.809	0.850
SVM	0.787	0.753	0.817	0.785
k-NN	0.758	0.638	0.758	0.757

VALUTAZIONE DEI RISULTATI 2

Al fine di stabilire quale o quali classificatori siano da considerarsi i migliori per tale problema, si inizia col filtrarli sulla base del valore dell'accuratezza di generalizzazione. In linea con l'obiettivo iniziale, vengono selezionati quegli algoritmi che superano la soglia di accuratezza dell'80%, quindi: Random forest e Naive bayes.

ALGORITMO	ACC. DI GENERALIZZAZIONE	SENS.	SPEC.	AUC
Random forest	0.835	0.774	0.806	0.886
Naive Bayes	0.826	0.783	0.861	0.897

Come secondo parametro di valutazione si considera l'AUC. Tuttavia, la differenza tra le due AUC non risulta tale da stabilire se un algoritmo sia migliore dell'altro. Si utilizza, pertanto, un t-test con $\alpha=0.05$ sulle differenze di AUC dei k-fold al fine di stabilire se vi è differenza significativa dei due valori. Per farlo si è assunto che le distribuzioni delle aree sotto la curva roc siano distribuite gaussianamente, ciò non ha permesso comunque di rifiutare l'ipotesi nulla che le medie siano uguali. Di conseguenza, si è effettuato il t-test anche sui valori di accuratezza, sensibilità e specificità.

Nemmeno i test su tali parametri hanno permesso di rifiutare l'ipotesi nulla, pertanto entrambi gli algoritmi possono essere considerati ugualmente validi a descrivere il problema.

Inoltre, l'analisi esplorativa della curva roc al variare del rapporto dei costi, non ha permesso di trarre conclusioni circa la scelta dell'algoritmo preferibile. Per entrambi, infatti, le prestazioni sono pressochè simili e la variazione del rapporto dei costi, e, quindi, della soglia decisionale, conduce agli stessi risultati in termini di sensibilità e di rate di falsi positivi.

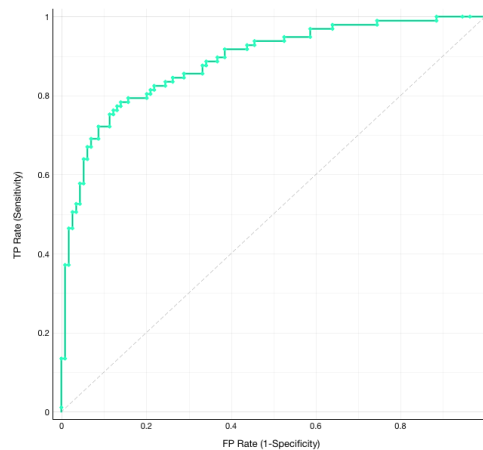


FIGURA 15:CURVA ROC NAIVE BAYES

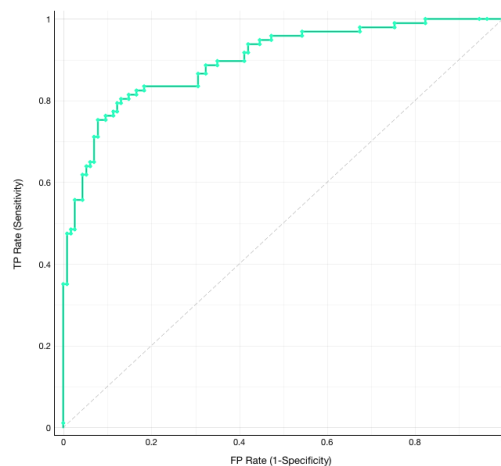


FIGURA 16:CURVA ROC RANDOM FOREST

Infine, non potendo scegliere uno dei due algoritmi sulla base della performance, si ritiene che la scelta migliore debba essere svolta sulla base della natura dell'utilizzatore finale del sistema di apprendimento automatico. Pertanto, viene selezionato il Naive Bayes in quanto è estremamente intuitivo dal punto di vista implementativo e fornisce ai medici uno strumento in grado di mostrare il contributo di ogni singola variabile sulla diagnosi. In particolare, verrà fornito l'algoritmo appreso su tutti i dati e non sui singoli fold, così che le sue prestazioni risultino migliori o uguali di quanto atteso.

Si riportano di seguito gli indici di qualità del classificatore selezionato corredati dell'intervallo di confidenza al 95% ottenuto da una distribuzione t-student a 9 gdl.

	INTERVALLO DI CONFIDENZA AL 95%
AUC	0.838 - 0.955
ACC. DI GENERALIZZAZIONE	0.778 – 0.875
SENSIBILITA'	0.684 – 0.882
SPECIFICITA'	0.809 – 0.913

DIFFUSIONE

L'algoritmo di apprendimento scelto potrebbe essere un notevole aiuto per i medici, se presentato sotto una forma intuitiva ed utilizzabile.

Una sua possibile implementazione consiste nel creare una semplice interfaccia grafica, fruibile su sistemi mobile o desktop, in cui sia possibile inserire i valori delle 9 variabili considerate tramite dei menu a tendina. A inserimento completato, il sistema fornisce la diagnosi con la relativa probabilità, con cui il medico può confrontarsi.

L'applicazione, una volta richiesta la diagnosi, invia automaticamente i valori inseriti al database regionale, nazionale e internazionale così da permettere all'algoritmo di essere aggiornato periodicamente con un numero di dati maggiore, portandolo così alle sue prestazioni ottimali. Inoltre, la presenza di database relativi a zone geografiche differenti consente di effettuare statistiche in maniera estremamente efficiente, data la loro facilità di interfacciarsi con i più diffusi sistemi informativi.

L'interfaccia permette anche di far variare la soglia decisionale sulla base del rapporto tra FP e FN, scelto dall'ASL competente o dal SSN, a seconda delle esigenze.

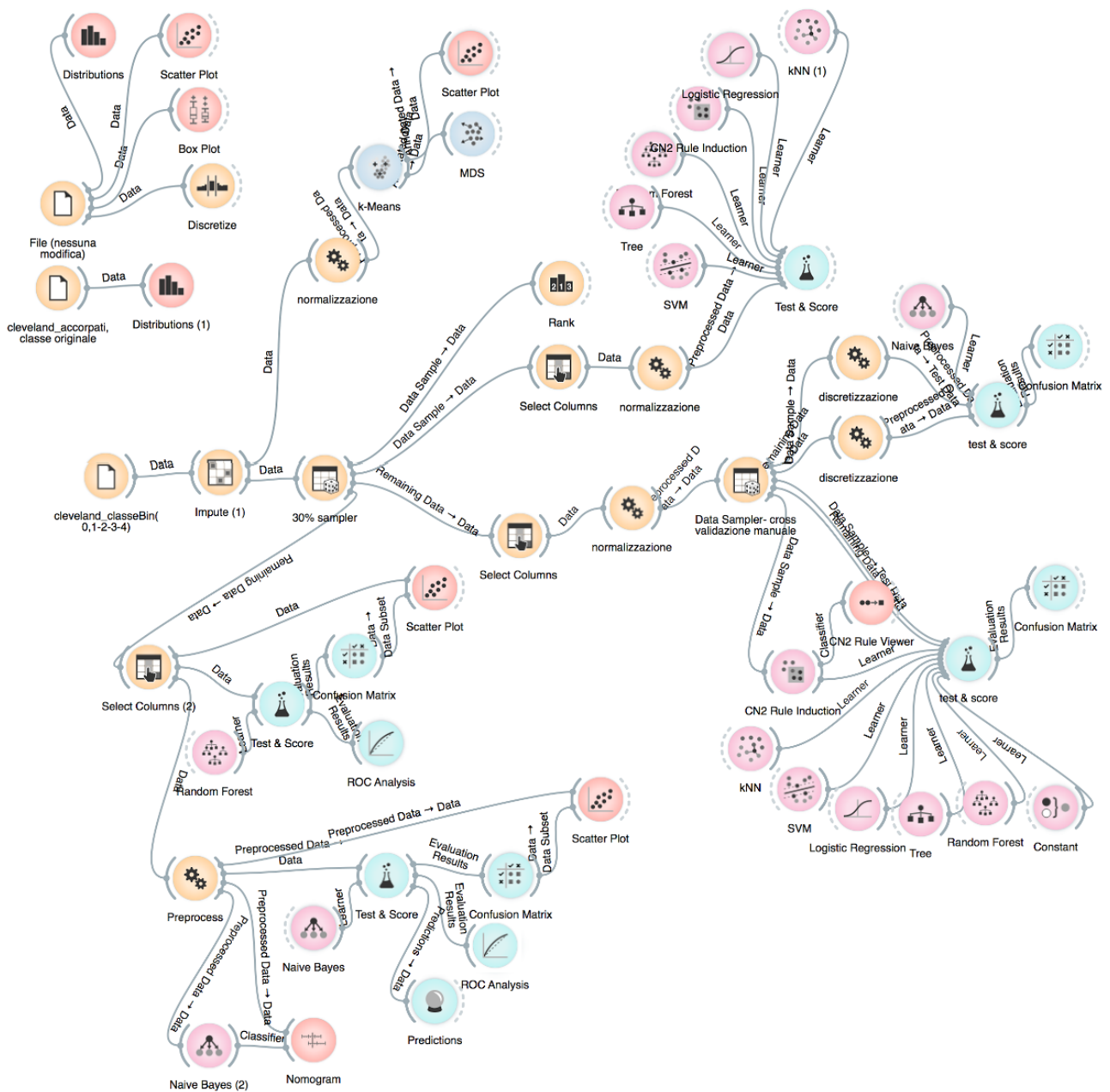
Gli operatori che usufruiscono di tale sistema sono preventivamente istruiti sul suo corretto utilizzo attraverso dei brevi corsi di formazione svolti nelle ore lavorative. Verranno, dunque, illustrati i modi di utilizzo e i benefici apportati da tale sistema in termini di efficienza ed efficacia diagnostica e costi.

Il sistema, così come è pensato, è utilizzabile solo dal personale specializzato, il medico cardiologo, poiché la presenza di alcune variabili, quali il *restecg*, *thal* o il *ca*, necessita di un esame specifico che non è possibile svolgere autonomamente. Si potrebbe, comunque, immaginare un'ulteriore scenario di implementazione dell'apprendimento automatico, sfruttabile da un comune utente. Ciò implicherebbe uno studio della performance degli algoritmi in assenza degli attributi relativi alla fluoroscopia e alla talassemia, che necessitano di un esame e del parere medico. Al contrario, i valori relativi all'elettrocardiogramma sono facilmente calcolabili tramite i moderni smartwatch.

L'interfaccia grafica risulta notevolmente semplificata rispetto a quella del medico e, ovviamente, non si può variare il rapporto tra i costi dei falsi positivi e i costi dei falsi negativi. L'utente avrebbe comunque un feedback della sua condizione e, sulla base dei risultati, potrebbe decidere di svolgere una visita dal cardiologo, favorendo la diagnosi precoce e riducendo, così, i costi del SSN.

APPENDICE 1

Progetto orange



APPENDICE 2

CODICE MATLAB

MUTUA INFORMAZIONE

```
%funzione che calcola la mutua informazione
function MI = mutuaInf(x1, x1Values, x2, x2Values)
MI = 0;
for i=1:length(x1Values)
    px1 = length(find(x1==x1Values(i)))/length(x1);
    for j=1:length(x2Values)
        px2 = length(find(x2==x2Values(j)))/length(x2);
        px1x2 = length(find(x1==x1Values(i) & x2==x2Values(j)))/length(x2);
        if(~isnan(px1x2*log2((px1x2)/(px1*px2))))
            MI = MI + px1x2*log2((px1x2)/(px1*px2))
        end
    end
end
end
end
```

FUNZIONE PER DISTRIBUZIONE DI PROBABILITÀ CUMULATIVA CONDIZIONATA DALL'ATTRIBUTO

```
% funzione che calcola la probabilità cumulativa di un dato valore della
% classe rispetto ad un attributo.
function cumuCond = cumulativaCondizionale(attributo, classe, valoreClasse)
    attributoOrdinato = sort(attributo);
    for(i=1:length(attributoOrdinato))
        prova = classe(find(attributo==attributoOrdinato(i)));
        provaFreq=prova(find(prova==valoreClasse));
        probCu(i) = (length(provaFreq))/length(prova);
    end
    probCu
    cumuCond = probCu;
end
```