Politecnico di Milano

FACOLTÀ DI INGEGNERIA DEI SISTEMI
Corso di Laurea Magistrale in Ingegneria Matematica

Tesi di laurea magistrale

Modelli statistici per la previsione del rischio cardiovascolare: analisi di segnali elettrocardiografici.

Relatore:
Prof. Anna Maria Paganoni

Correlatore:
Dott. Nicholas Tarabelloni

Candidato:
Silvia Giussani
Matricola 796076

Anno Accademico 2013–2014

Indice

Elenco delle figure

Elenco delle tabelle

Listings

Sommario

Le malattie cardiovascolari ischemiche sono una delle principali cause di decesso in tutto il mondo. In questa classe di patologie una diagnosi rapida è essenziale per una buona prognosi nel trattamento. In particolare, una procedura di classificazione automatica basata sull'analisi statistica di tracciati elettrocardiografici (ECG) sarebbe molto utile per permettere una diagnosi precoce. L'obiettivo di questa tesi consiste nello sviluppo e nell'applicazione di un algoritmo che consenta di fornire una diagnosi semi-automatica e in tempo reale della patologia del paziente, basandosi solo sulla morfologia dell'elettrocardiogramma e su altre covariate osservabili del paziente. Tale procedura rappresenta un valido cruscotto di supporto alle decisioni cliniche. Il dataset in esame consiste di 3,068 tracciati elettrocardiografici, sia di soggetti sani che di pazienti affetti da malattie cardiovascolari, raccolti utilizzando dispositivi di registrazione del segnale di Mortara Rangoni Europe s.r.l. e successivamente inviati con trasmissioni GSM al centro di raccolta del 118 di Milano. Come primo passo ci si occupa del pre-processing dei dati, rielaborandoli al fine di rendere possibile le sucessive analisi statistiche. Si procede poi con tecniche di clustering, utilizzando oppotuni modelli di regressione logistica. Quindi, è stata proposta una procedura di diagnosi automatica dei tracciati elettrocardiografici basata su un albero di classificazione. Infine, è stata sviluppata un'applicazione web, basata su *Shiny* [1], con lo scopo di rendere possibile l'utilizzo e la diffusione dell'algoritmo implementato presso un pubblico di non specialisti che, pur senza conoscere i dettagli statistici e implementativi del metodo proposto, possano facilmente utilizzarlo nella pratica quotidiana.

Key words: Segnale elettrocardiografico (ECG); Regressione logistica; Classificazione; Blocco di branca sinistra (LBBB); Blocco di branca destra (RBBB); Shiny (RStudio).

Abstract

Cardiovascular ischaemic diseases are one of the main causes of death all over the world. In this class of pathologies a quick diagnosis is essential for a good prognosis in the treatment. In particular, an automatic classification procedure based on statistical analysis of electrocardiographic traces (ECG) would be very helpful for an early diagnosis. The challenge of this work consists of developing and applying a procedure which enables semi-automatic and realtime diagnosis of the patients' disease based only on the morphology of ECG traces and on other observable variables. This procedure represents a valid support tool for clinical evaluations. The considered dataset consist of 3,068 electrocardiographic traces, either physiological or pathological, of patients whose ECG has been collected using Mortara Rangoni Europe s.r.l. instruments and sent to the 118 Dispatch Center in Milan by life-support personnel. Our attention is focused on a particular set of diseases: *Left Bundle Branch Block* (LBBB), *Right Bundle Branch Block* (RBBB) and *Atrial fibrillation*. The statistical analysis starts with a preprocessing step, where the dataset was prepared for the subsequent analysis. Then, a multivariate logistic regresion clustering procedure is carried out on the landmarks of the first lead of the ECG and on some other variables representing the patient's status. Hence, a new automatic diagnostic procedure, based on a classification tree, is proposed to classify ECG traces. Finally, a web application was created, using *Shiny* [1], with the purpose of making possible the use and the spread of the implemented algorithm to an audience of non-specialists, that without knowing the statistical and implementative details of the proposed method, can easily use it in everyday practice.

Key words: Electrocardiograph signal (ECG); Logistic regression; Clustering; Left Bundle Branch Block (LBBB); Right Bundle Branch Block (RBBB); Shiny (RStudio).

Ognuno è un genio.
Ma se si giudica un pesce
dalla sua abilità di
arrampicarsi sugli alberi,
lui passerà tutta la sua vita
a credersi stupido.

Albert Einstein.

Introduzione

Le applicazioni odierne della statistica riguardano dati sempre più complessi, in particolare per quanto concerne il settore biomedico: i macchinari necessari al monitoraggio di parametri vitali producono segnali, funzioni, immagini o una combinazione di questi. Sorge spontanea la ricerca di modelli per descrivere i fenomeni biologici in esame e di tecniche di inferenza statistica per riassumere la complessità di tali dati. Le malattie ischemiche cardiovascolari, quali l'infarto miocardico acuto, sono oggi una delle principali cause di morte in tutto il mondo. In Italia, sono responsabili del 44% della mortalità generale e della maggior parte delle emergenze sanitarie. Quasi tutte le chiamate che pervengono al 118 e che richiedono operazioni di soccorso, riguardano problemi cardiovascolari [2]. Nel caso di malattia ischemica coronarica, una diagnosi rapida è essenziale affinché il paziente abbia buone possibilità di recupero. Uno degli obiettivi di questa tesi è quindi quello di cercare di sfruttare i dati prodotti da dispositivi per la registrazione di segnali ECG presenti sulle auto mediche del primo soccorso, per costruire un algoritmo in grado di produrre una diagnosi precoce che possa garantire l'indirizzamento del paziente nella giusta struttura ospedaliera. In particolare l'obiettivo è di prevedere lo stato di salute di un paziente per mezzo di metodi tradizionali, come la classificazione supervisionata. Più nel dettaglio è stata utilizzata la regressione logistica a step successivi per poter stimare le probabilità di essere sano o affetto da una determinata patologia sulla base di alcuni valori di sintesi dell'elettrocardiogramma.

Nel Capitolo 1, dopo alcune premesse fisiologiche, si descrivono il dataset a disposizione e il pre-processing fatto su di esso per poter procedere con le analisi.
Nel Capitolo 2 si procede con un'analisi descrittiva del dataset. Vengono riportate le medie e le varienze delle variabili principali per ogni classe di pazienti. A partire dalle differenze emerse dai boxplot, si prospettano le linee guida per i modelli di classificazione, descritti nel dettaglio nel capitolo successivo.
Nel Capitolo 3 si descrive in generale la regressione logistica, il modello e i metodi di classificazione che da essa discendono.
Nel Capitolo 4 vengono riportati i risultati ottenuti applicando i sopradetti

metodi al dataset a disposizione, evidenziando le percentuali di misclassificazione.
Il Capitolo 5 è dedicato allo sviluppo di un'applicazione web, cioè un'interfaccia grafica che permetta all'utente (ad esempio personale clinico) di accedere all'utilizzo dei metodi proposti senza occuparsi o conoscere i dettagli computazionali e implementativi.

Tutte le analisi sono state effettuate utilizzando il software R, versione 3.1.1 [3].

Capitolo 1

Descrizione e preprocessing del dataset

1.1 Premessa: qualche nozione di elettrofisiologia cardiaca

Il cuore è un organo cavo posto al centro della cavità toracica, costituito prevalentemente da tessuto muscolare striato e, per questo, in grado di contrarsi. È suddiviso in quattro cavità: due atri e due ventricoli, rispettivamente destri e sinistri. Ciascun atrio comunica con il corrispondente ventricolo grazie a delle valvole, mentre la parte destra dell'organo resta totalmente divisa da quella sinistra grazie ad un setto longitudinale. I due lati hanno funzioni molto diverse: la parte destra è responsabile della circolazione del sangue venoso, mentre la sinistra di quella del sangue arterioso. Il sangue riesce a raggiungere qualsiasi distretto corporeo grazie alla pressione che si crea a seguito della contrazione del cuore. Quest'ultima è diretta conseguenza della propagazione di un segnale elettrico di depolarizzazione che ha origine nel nodo senoatriale, un piccolo agglomerato di muscolo cardiaco con sede nella parte superiore laterale dell'atrio destro. Il segnale elettrico provoca, in primis, una contrazione atriale, continua il suo percorso tramite le vie internodali che si dipartono dal nodo senoatriale e raggiunge, in questo modo, il nodo atrioventricolare. Questo è il responsabile principale del ritardo che intercorre tra contrazione atriale e contrazione ventricolare. Il segnale prosegue, dunque, lungo il fascio di His, un insieme di fibre che si suddivide in due parti, la branca destra e la branca sinistra, che culminano nelle fibre del Purkinjie. Da qui, il segnale di depolarizzazione raggiunge i muscoli ventricolari, nei quali avrà poi luogo la ripolarizzazione. Una rappresentazione del sistema di conduzione del cuore è riportata in Figura 1.1, mentre in Figura 1.2 sono rappresentate le fasi di depolarizzazione e ripolarizzazione che costituiscono un battito cardiaco.

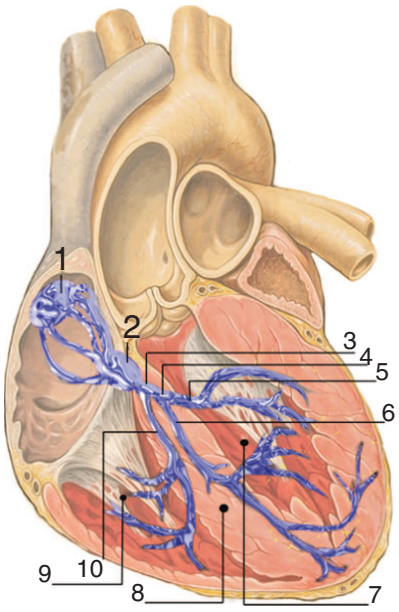


Figura 1.1: Sistema di conduzione del cuore: 1, nodo seno-ariale; 2, nodo atrio-ventricolare; 3, fascio di His; 4, branca sinistra; 5, fascio posteriore sinistro; 6, fascio anteriore sinistro; 7, ventricolo sinistro; 8, setto interventricolare; 9, ventricolo destro; 10, branca destra.

L'attività elettrica delle cellule cardiache determina un flusso di correnti all'interno del cuore e si manifesta con una variazione della differenza di potenziale sulla superficie della pelle che può essere misurata con opportuni apparati sperimentali. L'andamento in tempo di tali differenze di potenziale costituisce l'elettrocardiogramma (ECG). Da un punto di vista pratico, l'ECG è un insieme di 12 tracciati temporali di altrettante differenze di potenziale misurate tra punti differenti della superficie del corpo umano. Esso è la proiezione su 12 direzioni nello spazio tridimensionale del vettore cardiaco risultante dei momenti dei dipoli elettrici determinati nel cuore dall'avanzamento del fronte d'onda. Ognuna delle 12 derivazioni dell'ECG mostra il modulo del vettore cardiaco (e quindi l'attività elettrica) nella corrispondente direzione per ogni istante di tempo. Le 12 direzioni sono scelte in maniera tale da individuare una suddivisione dello spazio che esprima l'attività nelle orientazioni destro-sinistro, superiore-inferiore, anteriore-posteriore del corpo. Si nota subito che usare le proiezioni lungo 12 direzioni per rappresentare un vettore in uno spazio tridimensionale si abbia un'evidente ridondanza e correlazione nei tracciati delle derivazioni che rappresenta un interessante oggetto di studio. La disposizione di base degli elettrodi consta nella creazione di un triangolo con vertici idealmente collocati in prossimità delle radici degli arti superiori e della regione pubica.

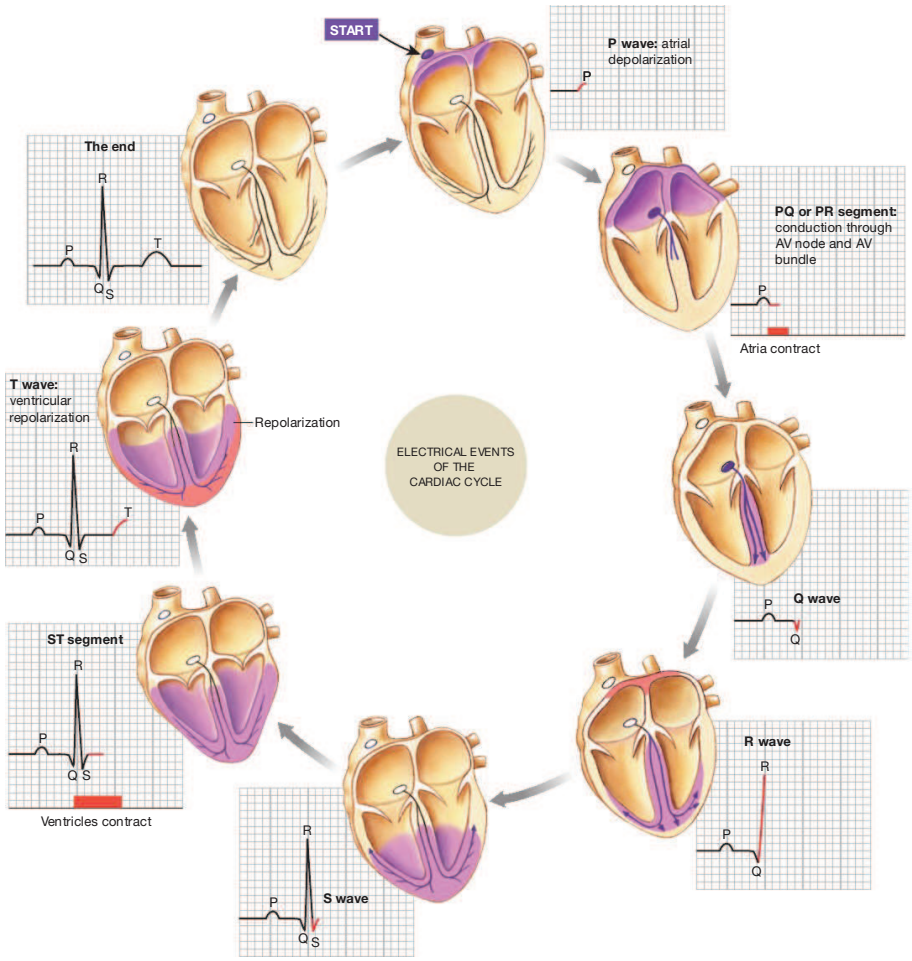


Figura 1.2: Correlazione tra un ECG e gli eventi elettrici nel cuore. Ad ogni fase del ciclo vengono evidenziate le regioni di depolarizzazione (in viola) e di ripolarizzazione (in arancione). Si veda [4].

Poiché la differenza di potenziale tra le radici e le estremità distali degli arti è trascurabile, per comodità di applicazione, si preferisce posizionare gli elettrodi sui polsi destro, sinistro e sulla caviglia sinistra.

Il triangolo che si forma viene comunemente chiamato **Triangolo di Einthoven** ed è riportato in Figura 1.3.
Le derivazioni standard sono dodici e si suddividono in tre tipi: bipolari, unipolari aumentate di Goldberger e unipolari precordiali di Wilson.

Derivazioni bipolari:
Sono tre derivazioni misurate su un piano frontale, in particolare sono:

Derivazione I: data dalla differenza tra il potenziale del polso sinistro ed il potenziale del polso destro;

Derivazione II: data dalla differenza tra il potenziale della caviglia sinistra ed il potenziale del polso destro;

Derivazione III: data dalla differenza tra il potenziale della caviglia sinistra ed il potenziale del polso sinistro.

È importante notare che queste derivazioni sono linearmente dipendenti, in particolare si ha che, per la legge di Kirchhof, I + III = II.

Derivazioni unipolari aumentate di Goldberger:
Anch'esse sono tre derivazioni misurate su un piano frontale, lungo le tre bisettrici del triangolo di Einthoven. In particolare misurano la differenza di potenziale tra uno dei vertici ed un punto di riferimento all'intersezione tra due resistenze identiche collegate rispettivamente ai due restanti vertici del triangolo stesso. L'aggettivo aumentate indica la necessità di amplificazione del segnale per poterlo rapportare alle altre derivazioni. Sono:

- Derivazione aVL;
- Derivazione aVR;
- Derivazione aVF.

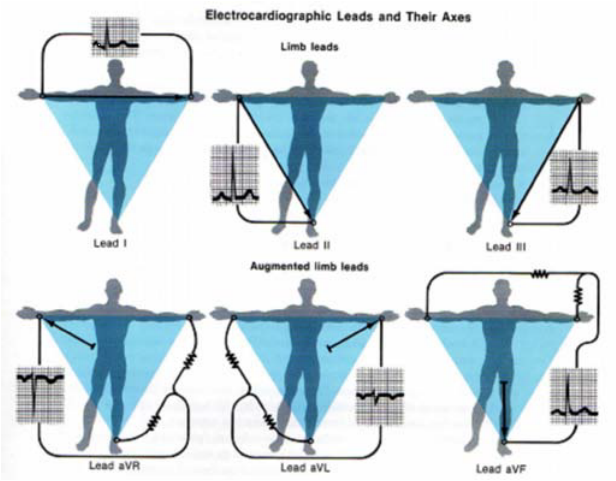


Figura 1.3: Differenza di potenziale delle derivazioni degli arti e triangolo di Einthoven

Derivazioni unipolari precordiali di Wilson:
Per ottenere quest'ultima tipologia di derivazioni è necessario un punto di riferimento che coincide con il centro del triangolo di Einthoven, ovvero il punto all'intersezione di tre resistenze identiche connesse ai tre vertici del triangolo stesso. Le differenze di potenziale che vengono registrate si riferiscono alla differenza di voltaggio tra il suddetto punto ed altri sei punti opportunamente collocati negli spazi intercostali, come in Figura 1.4. Si ottengono, dunque, sei derivazioni su un piano trasversale. Sono:

- Derivazione V1;
- Derivazione V2;
- Derivazione V3;
- Derivazione V4;
- Derivazione V5;
- Derivazione V6.

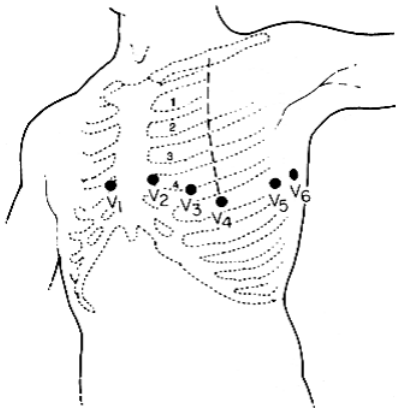


Figura 1.4: Posizione degli elettrodi per le derivazioni precordiali

Come già visto per le prime tre, esistono dipendenza e correlazione tra le derivazioni. Questo è intuitivamente giustificato dal fatto che si sta guardando allo stesso fenomeno da diversi punti di vista. Il segnale elettrico, dunque, si propaga tra le cellule cardiache che si depolarizzano e successivamente si ripolarizzano. Per ogni ulteriore dettaglio si veda [5]. La registrazione del segnale per mezzo di un tracciato elettrocardiografico (ECG) permette di ottenere un segnale per ciascuna derivazione. Si riporta in Figura 1.5 la rappresentazione stilizzata della derivazione I in condizioni normali, con supporto graduato che permette di identificare facilmente la durata dei tratti caratteristici, la cui variabilità può essere un indice di anormalità dell'ECG.

Ogni tratto di tale curva rappresenta una particolare fase del battito cardiaco. Il segnale, infatti, può essere partizionato in sottointervalli usando i landmarks P, Q, R, S, T ed eventualmente U, punti riconoscibili nella maggior parte dei tracciati ed aventi un particolare significato fisiologico.

- **Onda P:** è la prima onda che si genera nel ciclo, e corrisponde alla depolarizzazione degli atri. È di piccole dimensioni, poiché la contrazione degli atri non è così potente. La sua durata varia tra i 60 e i 120 ms, l'ampiezza è uguale o inferiore ai 2.5 mV.
- **Complesso QRS:** si tratta di un insieme di tre onde che si susseguono l'una all'altra, e corrisponde alla depolarizzazione dei ventricoli. L'onda Q è negativa e di piccole dimensioni, e corrisponde alla depolarizzazione del setto interventricolare; la R è un picco molto alto positivo, e corrisponde alla depolarizzazione dell'apice del ventricolo sinistro; la S è un'onda negativa anch'essa di piccole dimensioni, e corrisponde alla depolarizzazione delle regioni basale e posteriore del ventricolo sinistro. La durata dell'intero complesso è compresa tra i 60 e 90 ms. In questo intervallo avviene anche la ripolarizzazione atriale che però non risulta visibile perché mascherata dalla depolarizzazione ventricolare.
- **Onda T:** rappresenta la ripolarizzazione dei ventricoli. Non sempre è identificabile, perché può anche essere di valore molto piccolo.
- **Onda U:** se presente è indice di una ripolarizzazione tardiva dei ventricoli.

Ognuno di questi tratti ha una durata specifica indicativa di uno stato di salute. Nel momento in cui si notano anomalie in termini di durata delle fasi di depolarizzazione o ripolarizzazione è possibile che siano in corso patologie.

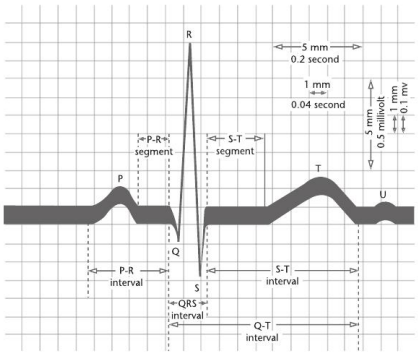
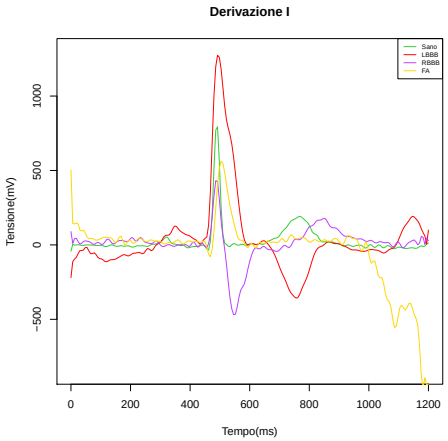


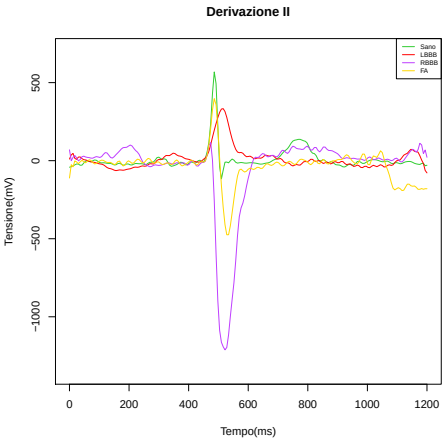
Figura 1.5: Tracciato della derivazione I in condizioni normali.

I dati a disposizione, illustrati nel dettaglio in seguito, riguardano in parte soggetti sani ed in parte pazienti affetti da una patologia cardiaca. Le più significative ai fini di questo lavoro sono: **LBBB**(Left Bundle Branch Block o, comunemente, Blocco di Branca Sinistra), **RBBB**(Right Bundle Branch Block o, comunemente, Blocco di Branca Destra) e **Fibrillazione Atriale**.
I blocchi di branca corrispondono ad un'anomalia di conduzione intraventricolare, più precisamente una mancata sincronizzazione nella depolarizzazione dei due ventricoli, che comporta un ritardo della depolarizzazione di una parte del muscolo ventricolare. Il tempo supplementare necessario per la depolarizzazione di tutto il muscolo caridaco provoca un allargamento del complesso QRS.
Nel caso di RBBB viene ostacolata la conduzione lungo la branca destra e quindi la parete interventricolare si depolarizza a partire dal lato sinistro. Questo comporta una piccola onda R in derivazione ventricolare destra (V1) e una piccola onda Q in derivazione ventricolare sinistra (V6). È necessario più tempo del normale affinché l'onda raggiunga il ventricolo destro a causa del blocco della normale via di conduzione, di conseguenza il ventricolo destro si depolarizza dopo il sinistro. Questo si ripercuote nel tracciato con la comparsa di una seconda onda R, detta **R1**, in V1 e un'onda S larga e profonda in V6.
Al contrario, se si verifica un LBBB, il ventricolo sinistro si depolarizza con un certo ritardo rispetto a quello destro. Ciò che i medici sfruttano ai fini di una diagnosi di LBBB consta, ad esempio, nell'osservare un complesso QRS più ampio rispetto alla norma, l'assenza del picco R o del picco Q nella derivazione V1 o la biforcazione del picco R nella derivazione V6.
Nella fibrillazione atriale, gli impulsi elettrici che danno luogo alla contrazione degli atri si attivano in modo totalmente caotico e frammentario dando origine a fronti d'onda multipli e a contrazioni disorganizzate e frammentarie. Questa deviazione dalla fisiologia ordinaria è riscontrabile in un'alterazione del tracciato elettrocardiagrafico. In particolare non ci sono onde P sull'ECG, ma solamente una linea irregolare. Tuttavia, la perdita della contrazione atriale, l'irregolarità del battito e l'aumento della pressione di riempimento possono compromettere la funzione ventricolare sinistra, in maniera variabile, e la tolleranza agli sforzi può essere pertanto ridotta. La fibrillazione atriale, inoltre, è un importante fattore di rischio per lo stroke (ictus).

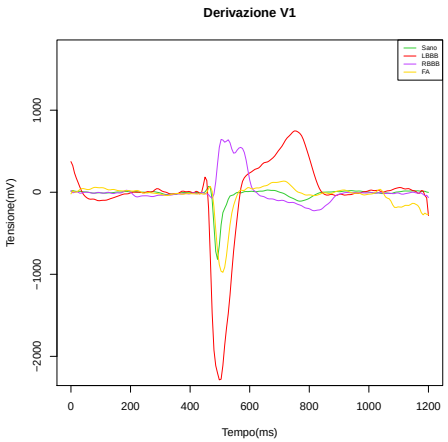
Si riportano in Figura 1.6 e in Figura 1.7 i tracciati delle 8 derivazioni standard, confrontando l'andamento delle diverse classi di pazienti.



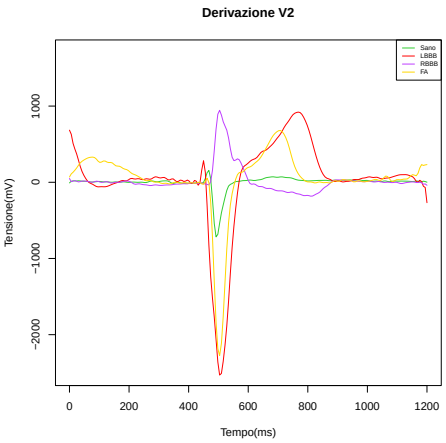
(a) Derivazione I a confronto



(b) Derivazione II a confronto

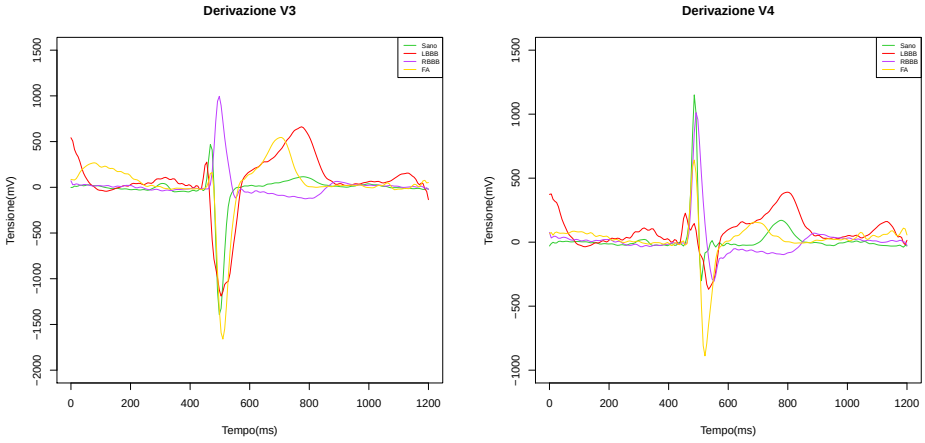


(c) Derivazione V1 a confronto



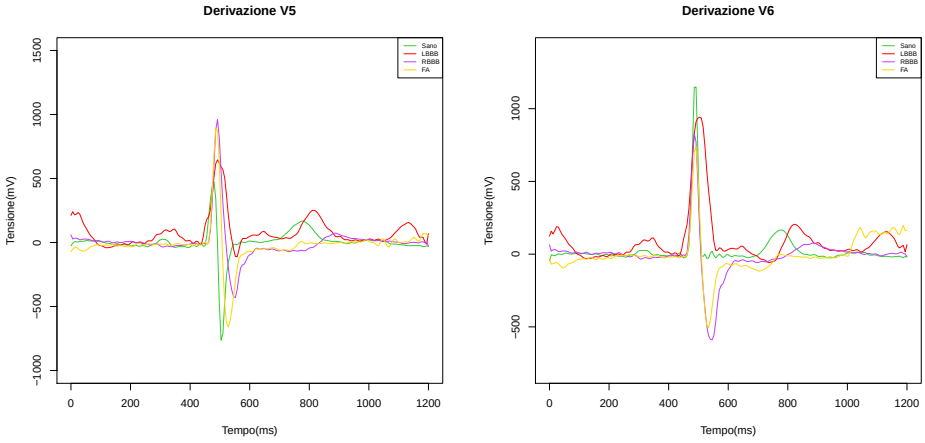
(d) Derivazione V2 a confronto

Figura 1.6: Derivazioni I, II, V1, V2. La linea verde corrisponde al tracciato di un soggetto sano, la linea rossa a quello di un paziente LBBB, la linea viola ad un paziente RBBB e infine la linea gialla al tracciato di un paziente affetto da fibrillazione atriale.



(a) Derivazione V3 a confronto

(b) Derivazione V4 a confronto



(c) Derivazione V5 a confronto

(d) Derivazione V6 a confronto

Figura 1.7: Derivazioni V3, V4, V5, V6. La linea verde corrisponde al tracciato di un soggetto sano, la linea rossa a quello di un paziente LBBB, la linea viola ad un paziente RBBB e infine la linea gialla al tracciato di un paziente affetto da fibrillazione atriale.

1.2 Preprocessing del dataset

Il dataset utilizzato proviene dai dati raccolti nell'ambito del progetto PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero), una collaborazione iniziata nel 2008 tra Azienda Regionale Emergenza Urgenza (AREU), Abbott Vascular and Mortara-Rangoni Europe s.r.l.[1] con l'intento di diffondere l'uso intensivo degli ECG come strumento diagnostico pre-ospedaliero e di costruire un nuovo database di ECG dalle caratteristiche mai registrate prima in altre banche dati sui disturbi cardiaci.
Sfruttando questa collaborazione, sono stati instalalti su tutte le Basic Rescue Units (BRU) dell'area urbana Milanese degli apparecchi per acquisire, registrare e trasferire via GSM gli ECG dei cittadini soccorsi dalle unità del 118, indipendentemente dai sintomi.
Il dataset consiste di $N = 6,734$ soggetti, dei quali 1,633 sono sani e 5,101 sono affetti da una patologia cardiaca. Come già discusso in precedenza, per questo lavoro sono stati considerati solo i pazienti affetti da **LBBB, RBBB** e **Fibrillazione atriale**. La numerosità del campione si riduce quindi a 3,068 pazienti di cui 1,633 sani, 420 RBBB, 314 LBBB e 701 fibrillazioni atriali.

Per ogni paziente sono a disposizione, in particolare, i tracciati di 8 derivazioni standard (sulle 12 totali per le ragioni di collinearità già menzionate nel Paragrafo 1.1), I, II, V1, V2, V3, V4, V5, V6, che può essere modellizzato statisticamente come un dato funzionale multivariato, e 12 variabili, che consistono in 8 landmarks e 4 variabili descrittive del paziente (Età, Sesso, *Firstline iterp* e *Interp text*), si veda Tabella 1.1. In questa analisi ci focalizeremo prevalentemente sull'analisi multivariata di alcuni parametri di sintesi del tracciato elettrocardiografico.
Ogni file del dataset, corrispondente ad un soggetto, è associato ad altri tre file. Il primo è chiamato *Text* e contiene informazioni tecniche, che possono essere utili per l'analisi del segnale, come i dettagli tecnici di registrazione di ogni derivazione, il tempo di onset e di offset dei più importanti sottointervalli dell'ECG e una diagnosi automatica effettuata con l'algoritmo Mortara-Rangoni VERITAS^{TM} e successivamente refertata da un medico. L'utilizzo di queste informazioni viene trattato nel dettaglio in seguito.
Il secondo file è chiamato *Rhythm* e contiene l'output della registrazione elettrocardiografica, cioè il segnale vero e proprio. Più precisamente registra 10 s (che corrispondono a 10,000 istanti di campionamento) del segnale ECG.
L'ultimo file, *Median*, è costruito a partire dal file *Rhythm* e costituisce un "battito di riferimento" della durata di 1.2 s (1,200 istanti di campionamento).

[1] Si ringrazia Mortara-Rangoni Europe s.r.l. per la gentile concessione dei dati.

Nelle successive analisi sono stati utilizzati esclusivamente i dati contenuti del file *Median* per quanto riguarda la parte **funzionale**, e le indicazioni sui landmarks dei sottointervalli presenti nel file *Text* per le analisi **multivariate**.
A questo punto, per rendere più facile l'mportazione dei dati, sono stati creati 8 nuovi dataset (uno per ogni derivazione) costituiti da una matrice $N \times (P+T)$ dove P è il numero delle variabili a disposizione per ogni paziente (si veda la Tabella 1.1) e $T = 1,200$ sono gli istanti di campionamento del tracciato elettrocardiografico della specifica derivazione.

Variabile	**Descrizione**	**Unità di misura**
Age	Età del paziente	Anni
Sesso	Sesso del paziente	{M, F}
Firstline interp	Diagnosi "primaria"	—
Interp text	Diagnosi "secondarie"	—
POnset	Istante in cui inizia l'onda P	[ms]
POffset	Istante in cui finisce l'onda P	[ms]
QRSOnset	Istante in cui inizia il complesso QRS	[ms]
QRSOffset	Istante in cui finisce il complesso QRS	[ms]
TOffset	Istante in cui finisce l'onda T	[ms]

Tabella 1.1: Elenco delle variabili presenti nel dataset originale

I dati grezzi, inizialmente affetti da rumore e non completi, sono stati rielaborati al fine di rendere possibile le sucessive analisi statistiche. Particolare attenzione è stata dedicata alla standardizzazione delle diagnosi. Nel dataset originale, infatti, la diagnosi è presente in forma descrittiva (estratta dal referto del medico) e non univocamente determinata. I pazienti sono stati raggruppati, utilizzando le informazioni presenti in letteratura medica, in 11 categorie, come riportato in Tabella 1.2. Per procedere con questa nuova classificazione è stata considerata la diagnosi *principale*, cioè la prima patologia che il medico ha refertato per ogni paziente. Alcune delle diagnosi subordinate presenti nella varibile *Interp text*, sono state prese in considerazione per creare altre variabili utili per classificare i pazienti, come ad esempio un indicatore che infichi la presenza o meno di fibrillazione atriale.

Classe	Numero di pazienti
Blocco atrio-ventricolare	284
Blocco di branca destra	420
Blocco di branca sinistra	314
Blocco di branca sinistra incompleto	51
Bradicardia	313
Fibrillazione atriale	701
Sani	1633
ST sopraslivellato	128
ST sottoslivellato	420
Tachicardia	598
Altro	1872
Totale	6,743

Tabella 1.2: Categorie di diagnosi

Dopo aver fatto gli opportuni controlli sul dataset, si è proceduto a calcolare e introdurre nuove variabili quali:

- **Diagnosi:** variabile che sostituisce *Firstline interp* e *Interp text*. Le diagnosi così ragruppate e standardizzate sono riportate in Tabella 1.2.
- **Ampiezza OndaP** = POffset - POnset
- **Ampiezza QRS** = QRSOffset - QRSOnset
- **Ampiezza OndaT** = TOffset - QRSOffset
- **OndaPflag** $\in \{0,1\}$, indica la presenza o meno dell'onda P nel tracciato elettrocardiografico; sarà utile per discriminare i pazienti affetti da fibrillazione atriale dagli altri. Per maggiori dettagli si veda il Capitolo 2.
- **FAflag** $\in \{0,1\}$, indica la presenza o meno della fibrillazione atriale nelle diagnosi, anche secondarie. Tutti i pazienti che hanno fibrillazione atriale come diagnosi primaria, avranno questo flag pari a 1. Si noti che le variabili *OndaPflag* e *FAflag* sono correlate, perché una è sintomo dell'altra, in quanto la fibrillazione atriale consiste in una contrazione individuale e indipendete delle fibre muscolari dell'atrio che genera un'anomalia nel tracciato ECG: l'assenza dell'onda P, sostituita da un linea irregolare. Si veda [6].

Per il calcolo delle due variabili seguenti è stata necessaria l'analisi del dato funzionale della derivazione I:

- **Area OndaT** $\in \mathbb{R}$ contiene il valore dall'area, con segno, sottesa dall'onda T.
- **Inversione OndaT** $\in \{0,1\}$, è uguale a 1 se c'è inversione del sottointervallo relativo all'onda T, 0 altrimenti. L'inversione dell'onda T si presenta nel caso in cui l'area sottesa da questo tratto di tracciato (espressa dalla variabile *Area OndaT*) è negativa, in caso contrario non ci sarà inversione. Questa variabile è importante, perché potrebbe essere utile nel discriminare pazienti LBBB da pazienti RBBB, si veda[7].

Capitolo 2

Analisi descrittiva dei dati

Prima di procedere con analisi di classificazione, è stata condatta un'analisi descrittiva del dataset. Sono riportate in Tabelle 2.1 le medie e le deviazioni standard delle variabili più importanti, per ogni classe di pazienti.

SANI

Covariate	Media	S.D.
Age	56	17.69
POnset	293.97	20.89
POffset	407.34	20.56
Ampiezza OndaP	113.37	12.97
QRSOnset	454.81	6.12
QRSOffset	549.16	9.40
Ampiezza QRS	94.35	10.32
TOffset	835.63	29.54
Ampiezza OndaT	286.47	29.07

LBBB

Covariate	Media	S.D.
Age	78	11.46
POnset	186.68	103.72
POffset	275.65	147.66
Ampiezza OndaP	88.97	50.32
QRSOnset	433.61	15.89
QRSOffset	576.34	18.66
Ampiezza QRS	142.74	26.44
TOffset	847.45	61.56
Ampiezza OndaT	271.11	58.89

RBBB

Covariate	Media	S.D.
Age	75	13.86
POnset	177.82	119.07
POffset	259.80	168.79
Ampiezza OndaP	81.99	54.90
QRSOnset	434.14	23.45
QRSOffset	575.57	23.98
Ampiezza QRS	141.43	21.74
TOffset	841.22	59.90
Ampiezza OndaT	265.65	58.36

FA

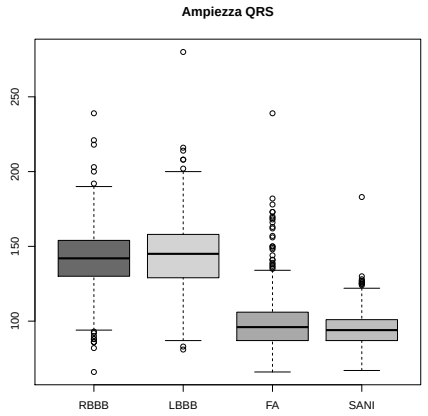
Covariate	Media	S.D.
Age	75	13.32
POnset	1.00	18.76
POffset	1.23	23.09
Ampiezza OndaP	0.00	4.34
QRSOnset	453.35	11.59
QRSOffset	552.26	13.58
Ampiezza QRS	98.90	17.83
TOffset	792.80	51.89
Ampiezza OndaT	240.54	51.02

Tabella 2.1: Medie e deviazioni standard delle variabili nelle quattro popolazioni

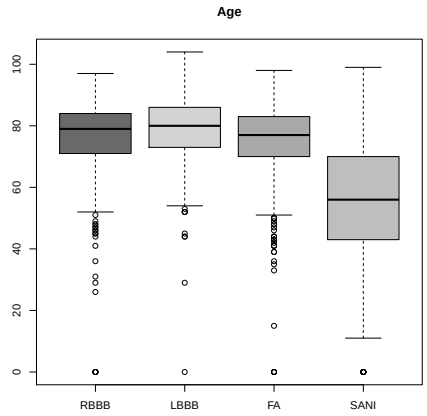
Da subito si possono fare alcune considerazioni:

- L'età è un fattore discriminante tra sani e malati;
- Nei pazienti affetti da BBB si riscontra un allargamento del complesso QRS;
- L'onda P risulta essere assente, come già evidenziato dall letteratura biomedica, nei pazienti con Fibrillazione Atriale.

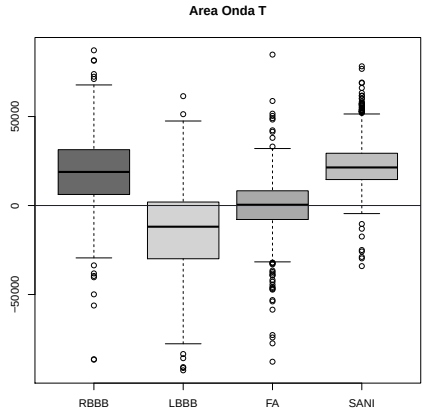
Per mostrare più facilmente queste differenze si riportano alcuni boxplot significativi in Figura 2.1.



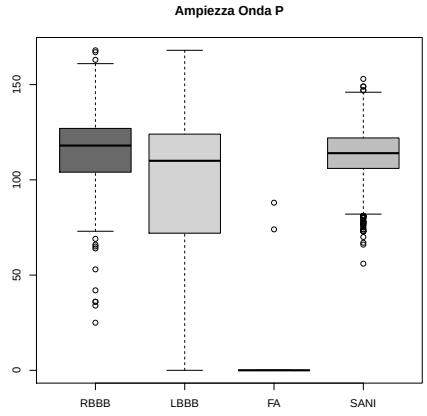
(a) Boxplot ampiezza QRS



(b) Boxplot Età



(c) Boxplot area OndaT



(d) Boxplot ampiezza Onda P

Figura 2.1: Boxplot di variabili rilevanti

Ora, al fine di quantificare e validare le differenze tra i gruppi emerse dai boxplot di Figura 2.1, si procede con una fase di test che hanno lo scopo di confermare cio che siè evinto via metodo grafico.

2.1 Test di Kruskal-Wallis

Il test di Kruskal-Wallis è un metodo non parametrico per verificare l'uguaglianza delle mediane di diversi gruppi, usato cioè per verificare che tali gruppi provengano da una stessa popolazione (o da popolazioni con uguale mediana). Questo metodo è il corrispondente non parametrico dell'analisi di varianza (ANOVA) in cui i dati vengono sostituiti dal loro rango, e viene solitamente usato quando non può essere assunta una distribuzione normale della popolazione.
L'ipotesi nulla consiste nell'affermare che non ci siano differenze tra le mediane, e quindi che i k gruppi possano essere considerati un campione proveniente da un'unica popolazione. L'ipotesi alternativa asserisce che almeno due degli effetti dei gruppi non sono uguali. Per ulteriori informazioni si faccia riferimento a [8].

2.1.1 Notazione

Si considera un dataset di $M = \sum_{j=1}^{k} m_j$ osservazioni, con m_j osservazioni del j-esimo gruppo, j=1,...,k.

Gruppi			
1	2	...	k
X_{11}	X_{12}	...	X_{1k}
X_{21}	X_{22}	...	X_{2k}
$\vdots$	$\vdots$		$\vdots$
$X_{m_1 1}$	$X_{m_2 2}$	...	$X_{m_k k}$

Tabella 2.2: Divisione in gruppi del dataset

2.1.2 Postulati

1. Le N variabili aleatorie $\{X_{1j}, X_{2j}, \ldots, X_{m_j j}\}$, j $\in \{1, \ldots, k\}$, sono mutuamente indipendenti;

2. Per ogni j $\in \{1, \ldots, k\}$ fissato, le m_j variabili aleatorie $\{X_{1j}, X_{2j}, \ldots, X_{m_j j}\}$ sono un campione aleatorio proveniente dalla distribuzione continua F_j;

3. Le distribuzioni $F_1, \ldots, F_k$ sono legate dalla relazione

$$F_j(t) = F(t - \tau_j) \qquad t \in \mathbb{R} \quad j \in \{1, \ldots, k\} \tag{2.1}$$

dove F è una funzione di densità continua con media incognita θ e con τ_j l'effetto aleatorio per il j-esimo gruppo.

Le ipotesi (1), ..., (3) sono equivalenti alla rappresentazione:

$$X_{ij} = \theta + \tau_j + e_{ij} \qquad i = \{1, \ldots, m_j\},\ j = \{1, \ldots, k\} \tag{2.2}$$

dove θ è la mediana totale, τ_j è l'effetto del j-esimo gruppo e e_{ij} costituisce un campione aleatorio con mediana nulla. Si noti che sotto l'ipotesi di normalità, le mediane (θ e 0) coincidono con le rispettive medie.

2.1.3 Ipotesi del test

$$H_0 : [\tau_1 = \ldots = \tau_k] \tag{2.3}$$

Quest'ipotesi nulla asserisce che tutte le distribuzioni F_k nell'equazione (2.1) sono uguali, cioè $F_1 \equiv F_2 \equiv \ldots \equiv F_k \equiv F$.

$$H_1 : [\tau_1, \ldots, \tau_k \text{ non tutti uguali}] \tag{2.4}$$

L'ipotesi alternativa invece asserisce che almeno due degli affetti dei gruppi siano diversi.

2.1.4 Metodo

Per calcolare la statistica di Kruskal-Wallis, H, per prima cosa si devono ordinare le M osservazioni dei k gruppi in ordine crescente. Sia r_{ij} il rango di X_{ij}, si definiscono le seguenti quantità:

$$R_j = \sum_{i=1}^{m_j} r_{ij} \qquad R_{.j} = \frac{R_j}{m_j},\ j = 1, \ldots, k. \tag{2.5}$$

Quindi, a titolo di esempio, R_1 è la somma dei ranghi congiunti del gruppo 1, mentre $R_{.1}$ è il rango medio delle stesse osservazioni. La statistica di Kruskal-Wallis, H, è data da:

$$H = \frac{12}{M(M+1)} \sum_{j=1}^{k} m_j \left(R_{.j} - \frac{M+1}{2} \right)^2 = \left(\frac{12}{M(M+1)} \sum_{j=1}^{k} \frac{R_j^2}{m_j} \right) - 3(M+1) \tag{2.6}$$

dove (M+1)/2 è il rango medio assegnato nell'ordinamento congiunto.
Per testare H_0 contro l'ipotesi alternativa H_1 al livello α si:

$$\text{Rifiuta } H_0 \text{ se} \quad H \geq h_\alpha \tag{2.7}$$

dove h_α, il quantile di livello $1-\alpha$ di H, è tale che la probabilità dell'errore di primo tipo sia pari ad α.

2.1.5 Approssimazione asintotica

Quando H_0 è vera e il $\min(m_1, \ldots, m_k) \to \infty$, la statistica H ha una distribuzione asintototica χ^2 con (k-1) gradi di libertà.
L'approssimazione χ^2 per la procedura (2.7) è

$$\text{Rifiuta } H_0 \text{ se} \quad H \geq \chi^2_{k-1,\alpha} \tag{2.8}$$

Dove $\chi^2_{k-1,\alpha}$ è il quantile di ordine $1-\alpha$ di una $\chi^2(k-1)$.

2.2 Test di Levene

Il test di Levene è utilizzato per fare inferenza sulla varianza di k gruppi. Molti test statistici assumono l'uguglianza delle varianze tra i gruppi; il test di Levene può essere utilizzato per verificare quest'assunzione. Questo test è un'alternativa al **Bartlett test** quando non si può assumere la normalità dei dati.
Il test di Levene è spesso utilizzato prima di fare un confronto tra medie. Quando il test indica di rifiutare l'ipotesi di omoschedasticità dei dati, allora occorre adottare test per la media generalizzati, che non assumano l'ipotesi di omogeneità delle varianze. Questa però non è l'unica situazione in cui il test di Levene viene utilizzato, infatti è anche utile in sè proprio al fine di testare l'uguaglianza di più popolazioni. Si vedano [9] e [10].

2.2.1 Ipotesi del test

$$H_0 : [\sigma_1^2 = \ldots = \sigma_k^2] \tag{2.9}$$

L'ipotesi (2.9) coincide con l'assunzione di **omoschedasticità**.

$$H_1 : [\sigma_1^2, \ldots, \sigma_k^2 \text{ non tutte uguali}] \tag{2.10}$$

2.2.2 Metodo

Utilizzando la stessa notazione introdotta nel Paragrafo 2.1.1, la statistica di Levene W è data da:

$$W = \frac{(M-k)}{(k-1)} \frac{\sum_{j=1}^{k} m_j (Z_{.j} - Z_{..})^2}{\sum_{j=1}^{k} \sum_{i=1}^{m_j} (Z_{ij} - Z_{.j})^2} \tag{2.11}$$

dove

- $Z_{ij} = \begin{cases} |X_{ij} - \bar{X}_j| & : \bar{X}_j \text{ è la media del j-esimo gruppo} \\ |X_{ij} - \tilde{X}_j| & : \tilde{X}_j \text{ è la mediana del j-esimo gruppo} \end{cases}$;
- $Z_{.j} = \frac{1}{m_j} \sum_{i=1}^{m_j} Z_{ij}$ è la media delle Z_{ij} nel j-esimo gruppo;
- $Z_{..} = \frac{1}{M} \sum_{j=1}^{k} \sum_{i=1}^{m_j} Z_{ij}$ è la media di tutte le Z_{ij}.

Le definizioni di Z_{ij} sono entrambe valide, nell'applicazione ai dati in esame è stata utilizzata la **mediana**.
Per testare H_0 contro l'ipotesi alternativa H_1 al livello α si

$$\text{Rifiuta } H_0 \text{ se} \quad W \geq F_{(\alpha,k-1,M-k)} \tag{2.12}$$

dove $F_{(\alpha,k-1,M-k)}$ è il quantile di ordine $1-\alpha$ della distribuzione F, con (k-1) e $(M-k)$ gradi di libertà.
Se il p-value del test di Levene è inferiore ad un certa soglia di significatività (spesso pari a 0.05), l'ipotesi nulla viene rifiutata e si può concludere che esiste una differenza significativa tra le varianze delle due popolazioni.

2.3 Applicazione ad un dataset di ECG

In questa sezione vengono applicati test descritti al paragrafo 2.1 e 2.2 ai dati in esame, per validare le differenze emerse dai boxplot e dell'analisi esplorativa delle medie e varianze.
Ciascun test è stato eseguito su 4 raggruppamenti dei dati, come riportato nella Tabella 2.4 e nella Tabella 2.3. Le analisi sono state svolte utilizzando il pacchetto R [11]. I *p-value* sono tutti inferiori alla soglia di significatività ($\bar{\alpha} = 0.05$) e questo avvalora le supposizioni che si faranno per procedere con la classificazione.

Sono riportati in Tabella 2.3 i p-value del test di Kruskal-Wallis applicato ai dati rappresentati in Figura 2.1.

Variabile	**Gruppi**	**P-value**
Ampiezza QRS	(LBBB + RBBB) vs. (FA + SANI)	$\ll 10^{-16}$
Age	(LBBB + RBBB + FA) vs. SANI	$\ll 10^{-16}$
Area onda T	(RBBB + FA+ SANI) vs. LBBB	$\ll 10^{-16}$
Ampiezza onda P	(LBBB + RBBB + SANI) vs. FA	$\ll 10^{-16}$

Tabella 2.3: P-value dei test di Kruskal-Wallis

Sono riportati in Tabella 2.4 i p-value del test di Levene applicato ai dati rappresentati in Figura 2.1.

Variabile	**Gruppi**	**P-value**
Ampiezza QRS	(LBBB + RBBB) vs. (FA + SANI)	$\ll 10^{-16}$
Age	(LBBB + RBBB + FA) vs. SANI	$\ll 10^{-16}$
Area onda T	(RBBB + FA+ SANI) vs. LBBB	$\ll 10^{-16}$
Ampiezza onda P	(LBBB + RBBB + SANI) vs. FA	$\ll 10^{-16}$

Tabella 2.4: P-value dei test di Levene

A partire da queste differenze, sia in media che in variabilità, espresse dai boxplot e validate dalla significatività dei p-value, si può procedere allo sviluppo di un modello di classificazione.

Come spiegato in precedenza l'obiettivo principale di questa Tesi è quello di riuscire a discriminare, utilizzando features multivariate, i soggetti in classi di patologia e quindi anche i pazienti **LBBB** dai pazienti **RBBB**. Considerando quanto emerge dall'analisi descrittiva dei dati, per fare ciò sembra che ci si possa avvalere delle seguenti considerazioni:

- La presenza o meno dell'**onda P** discrimina fra pazienti affetti da **fibrillazione atriale** da tutti gli altri;

- L'ampiezza del complesso **QRS** distingue fra **blochi di branca** e gli altri;
- L'inversione dell'**onda T** è peculiare dei pazienti **LBBB**; le Figure 2.2 e 2.3 ne mostrano alcuni esempi.

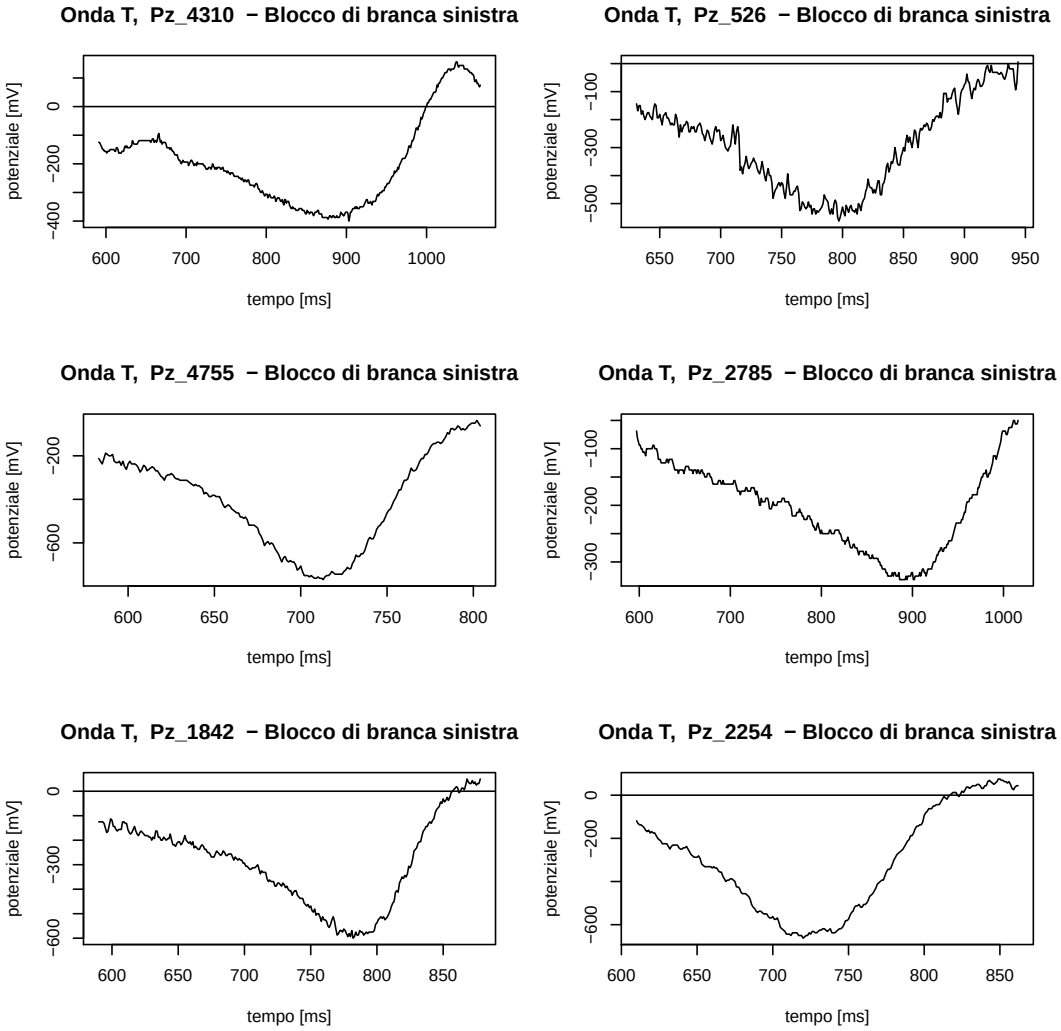


Figura 2.2: Esempi di inversione Onda T

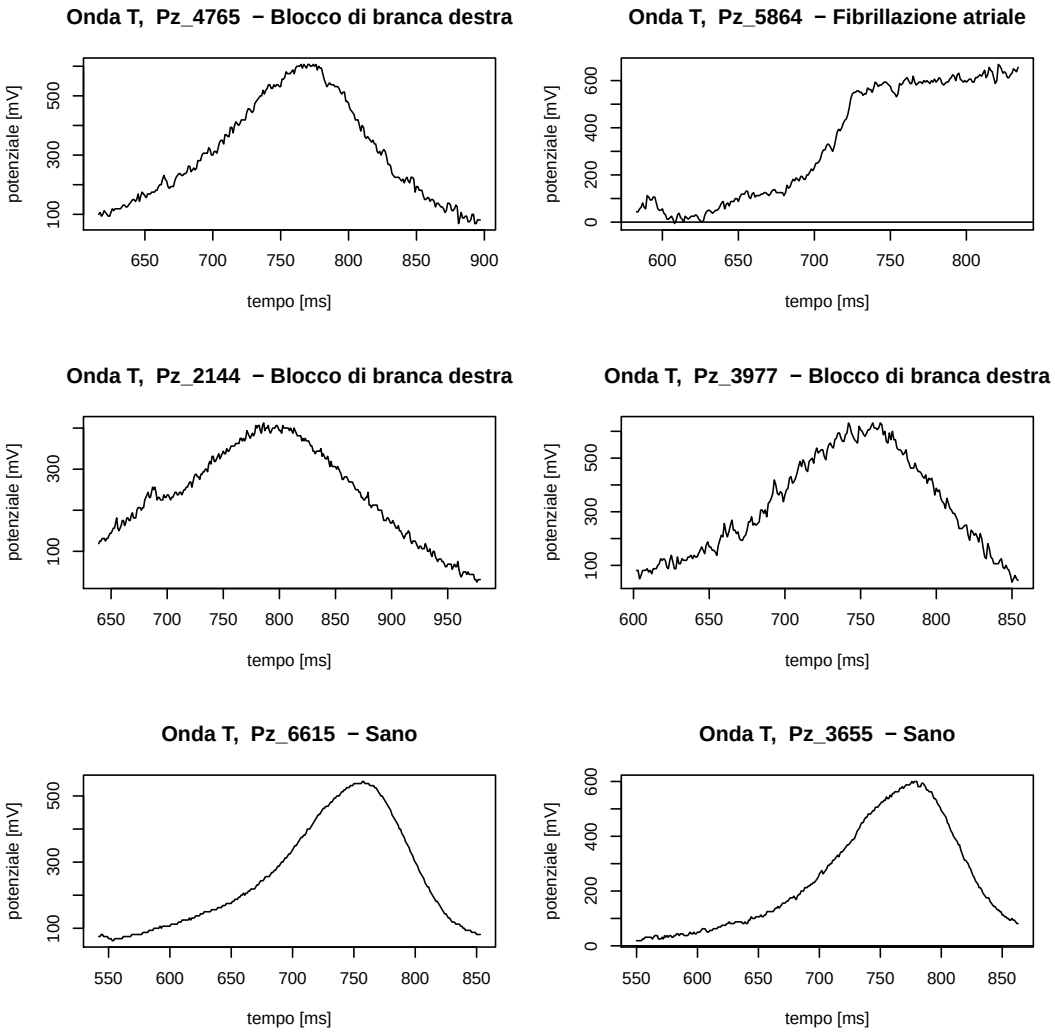


Figura 2.3: Esempi di onda T positiva

Capitolo 3

Richiami teorici sulla classificazione tramite regressione logistica

3.1 Modello

Come anticipato in precedenza, uno degli scopi di questa tesi è cercare di prevedere lo stato di salute o malattia di un paziente. Per far questo è stato adottato il modello di regressione logistica:

$$\boxed{logit(p_i) = \beta_0 + \sum_{i=1}^{p} \beta_i \cdot z_i} \tag{3.1}$$

dove :

- $logit(p_i) = \log \frac{p_i}{1-p_i}$;
- p_i è la probabilità che Y_i=1, dove $Y_i \sim Be(p_i)$ rappresenta il verificarsi della malattia;
- $Y_i \in \{0, 1\}$ indica lo stato di salute o malattia dell'i-esimo paziente;
- z_i è l'i-esimo regressore.

Date le convenzioni precedenti, si possono introdurre le quantità:

$$P(Y_1 = y_1, \ldots, Y_N = y_N) = \prod_{i=1}^{N} (p_i)^{y_i} (1 - pi)^{1-y_i} \tag{3.2}$$

$$p_i = P(Y_i = 1 \mid Z = z_i) = \Lambda(z_i \cdot \beta_i) = \frac{e^{\beta_0 + z_i \cdot \beta_i}}{1 + e^{\beta_0 + z_i \cdot \beta_i}} \tag{3.3}$$

che rappresentano, dato il modello, le probabilità che si verifichi un certo esito sperimentale. Esse permettono di costruire la cosiddetta **curva logistica**, che esprime la relazione non lineare fra p e il predittore z, mostrata in Figura 3.1.
Il valore di β_0 in (3.3) fornisce il valore $\frac{e^{\beta_0}}{1+e^{\beta_0}}$ quando $z = 0$.

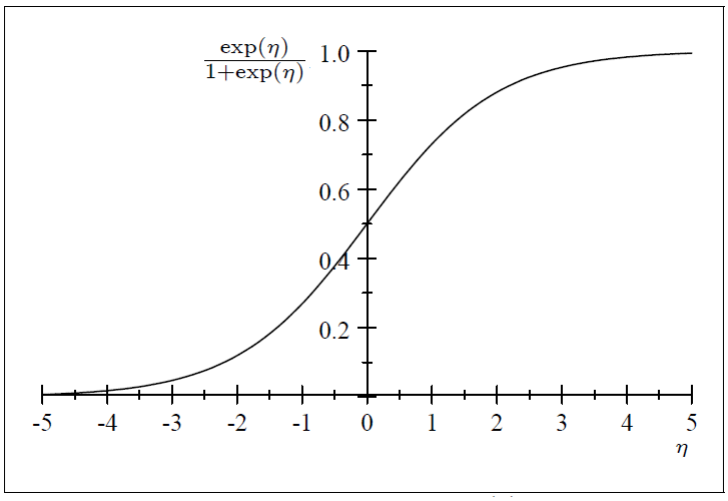


Figura 3.1: Curva logistica

I parametri β_i nella curva logistica, determinano quanto velocemente varia p al variare di Z, ma la loro interpretazione non è semplice come può essere quella dei parametri della regressione lineare, perchè in questo caso la relazione non è lineare né in Z né in β_i. Tuttavia è possibile sfruttare la relazione lineare per i log odds, infatti $\log \frac{p_i}{1-p_i} = \beta_0 + \sum_{i=1}^{p} \beta_i \cdot z_i$.
La stima dei parametri β può essere ottenuta tramite il metodo di **Massima verosimiglianza**.
La verosimiglianza L è data dalla probabilità congiunta della distribuzione, valutata nelle realizzazioni y_i. Si ottiene quindi

$$L(\beta_0, \beta_1, \ldots, \beta_p) = P(Y_1 = y_1, \ldots, Y_N = y_N) = \prod_{i=1}^{N} (p_i)^{y_i} (1 - pi)^{1-y_i} \quad (3.4)$$

dove p_i è definito nell'equazione (3.2).
I valori dei parametri che massimizzano la verosimiglianza non possono essere determinati come soluzioni in forma chiusa come nel caso teorico dei modelli lineari. Infatti essi sono determinati in modo numerico, a partire da un valore iniziale (*initial guess*) e iterando finché la precisione della stima risulta soddisfacente. Questa procedura prende il nome di **Iteratively reweighted least squares method**, si veda [12].
Denotiamo i valori di stima di massima verosimiglianza ottenuti numericamente con $\hat{\beta}$.

3.2 Classificazione

Una volta definito il modello e stimati i parametri $\hat{\beta}$ si può proseguire con la classificazione. La regressione logistica serve, ad esempio, per effettuare una classificazione supervisionata. A partire da un dataset noto, cioè formato da osservazioni di cui si conoscono sia i predittori sia l'etichetta che ne indica la classe di appartenenza (nel caso in esame, la patologia), si calcolano i coefficienti del modello. Poi, esaminando nuovi individui che si vogliono classificare sulla base della sola conoscenza dei predittori, si calcola la probabilità $\hat{p}$ per ogni individuo e lo si assegna alla classe che ha probabilità maggiore.
La classificazione ottenuta con la regressione logistica non si riduce semplicemente all'attribuzione di un'etichetta ad ogni unità statistica, ma in essa sono contenute informazioni più precise sul grado di sicurezza della classificazione stessa. Infatti per ogni unità statistica i si calcola $\hat{p}_i$, che inidca la probabilità di appartenere alla prima popolazione. A partire dalla probabilità $\hat{p}_i$ esistono tre modi per definire il classificatore:

1. Soglia fissata a 0.5 $\longrightarrow \hat{p}_i \geq 0.5$;
2. Soglia fissata alla media empirica $\longrightarrow \hat{p}_i \geq \bar{p}$;
3. Soglia fissata a partire dall'analisi della curva ROC, come si può vedere in Figura 3.2.

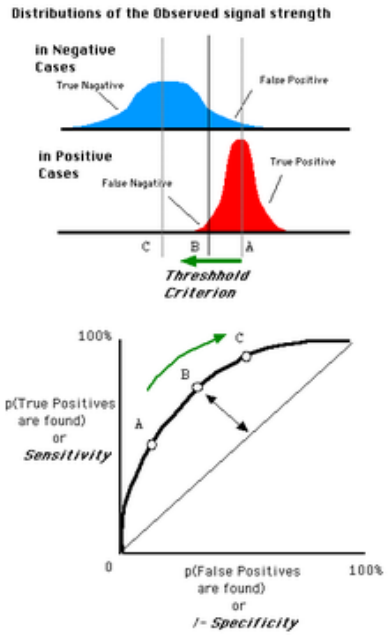


Figura 3.2: Andamento della soglia di un classificatore tramite curva ROC

3.3 Curva ROC

Per avere un'ulteriore visualizzazione grafica della regressione logistica, al variare della soglia, si può calcolare la curva ROC. La curva ROC (Receiver Operating Characteristic o Relative Operating Characteristic) è un grafico in cui vengono rappresentate le performances di un classificatore in termini di **False Positive Rate** e **True Positive Rate**. Si supponga da ora in avanti, per semplicità, un classificatore binario pensato per discriminare soggetti sani da soggetti malati. Siano:

- TN (True Negative) i soggetti sani classificati come sani;
- FP (False Positive) i soggetti sani classificati come malati;
- FN (False Negative) i soggetti malati classificati come sani;
- TP (True Positive) i soggetti malati classificati come malati;
- FPR (False Positive Rate) = 1 - Specificità;
- TPR (True Positive Rate) = Sensibilità.

	$\hat{0}$	$\hat{1}$
0	TN	FP
1	FN	TP

Tabella 3.1: Tabella di misclassificazione

- **Sensibilità**= $\frac{TP}{(TP+FN)}$;
- **Specificità**= $\frac{TN}{(TN+FP)}$.

La notazione utilizzata in Tabella 3.1 è la seguente:

- **0** indica un'unità statistica appartenente alla popolazione dei sani;
- **1** indica un'unità statistica appartenente alla popolazione dei malati;
- $\hat{\mathbf{0}}$ indica un'unità statistica classificata come appartenente alla popolazione dei sani;
- $\hat{\mathbf{1}}$ indica un'unità statistica classificata come appartenente alla popolazione dei malati.

La curva ROC è l'insieme delle coppie (FP, TP) al variare di un certo parametro del classificatore. In un classificatore a soglia, si calcolano la frazione di veri positivi e quella di falsi positivi per ogni possibile valore della soglia; tutti i punti così ottenuti nello spazio FP-TP descrivono la curva ROC.
La capacità discriminante di un classificatore, ossia la sua attitudine a separare propriamente la popolazione in sani o malati, è proporzionale all'estensione dell'area sottesa dalla curva ROC (**AUC, Area Under Curve**) ed equivale alla probabilità che il risultato di un test su un individuo estratto a caso dal gruppo dei malati sia superiore a quello di uno estratto dalla popolazione dei non malati. Nel caso di un test perfetto, ossia che non restituisce alcun falso positivo né falso negativo, la AUC passa attraverso il punto (0,1) e il suo valore corrisponde all'area dell'intero quadrato che ha come vertici (0,0), (0,1), (1,1), (1,0). Questo equivale ad una probabilità di classificazione corretta del 100%. Al contrario la ROC per un test privo di valore informativo corrisponde alla diagonale del quadrato sopracitato con corrispondente AUC=0.5. Questa situazione corrisponde al caso di un classificatore random, che assegna uguale probabilità di appartenenza alle due popolazioni.
Si noti che l'analisi della bontà di un classificatore effettuata tramite AUC assegna stessa importanza alla specificità e alla sensibilità, mentre in alcuni casi è necessario differenziare il peso da assegnare ai sopradetti parametri.
Si riporta in Tabella 3.2 la scala di interpretazione della capacità discriminante di un classificatore, [13]:

$AUC = 0.5$	classificatore non informativo
$0.5 < AUC \leq 0.7$	classificatore poco accurato
$0.7 < AUC \leq 0.9$	classificatore moderatamente accurato
$0.9 < AUC < 1$	classificatore altamente accurato
$AUC = 1$	classificatore perfetto

Tabella 3.2: Capacità discriminante di un classificatore

In conclusione si può affermare che l'utilizzo della curva ROC, come criterio di valutazione della bontà di un classificatore, sia più flessibile rispetto al criterio legato alla fissazione di una soglia, in quanto offre la possibilità di visualizzare il trade-off tra specificità e sensibilità.

Capitolo 4

Applicazione agli ECG

In questo capitolo si focalizza l'attenzione sull'applicazione dei modelli teorici descritti nel Capitolo 3 al caso reale in esame. Le differenze tra le diverse popolazioni di pazienti evidenziate dall'analisi descrittiva e in particolar modo dai boxplot di Figura 2.1, sono state tradotte in un **albero di classificazione**.
Per albero di classificazione si intende un processo decisionale gerarchico che, attraverso tre step successivi, riesce ad assegnare una diagnosi ad ogni elettrocardiogramma. Lo schema di classificazione è riassunto dal diagramma in Figura 4.1.

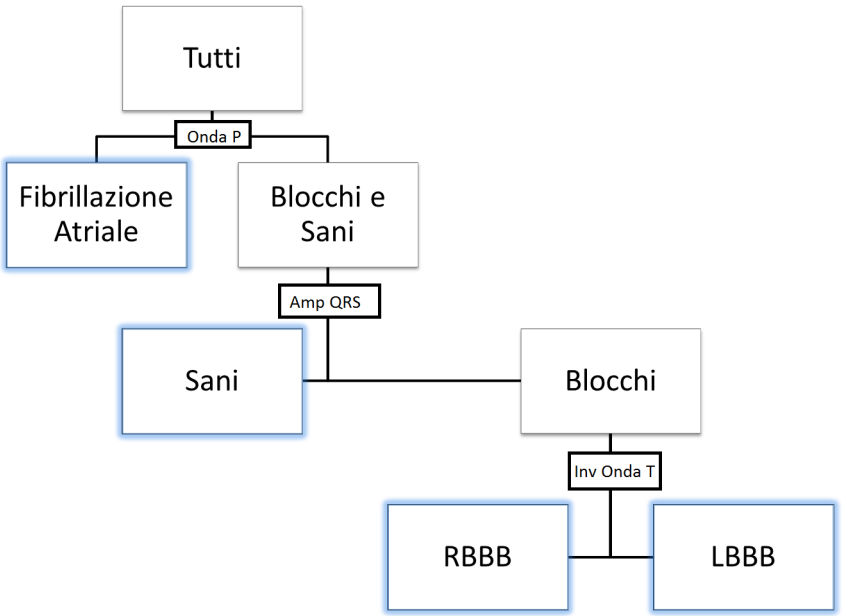


Figura 4.1: Schema dell'albero di classificazione

Il nuovo ipotetico paziente che entra nello studio e deve essere catalogato passa attraverso tre classificatori:

1. basato sulla presenza o meno dell'onda P;
2. basato sull'ampiezza del complesso QRS;
3. basato sulla presenza o meno dell'inversione dell'onda T.

Come prima cosa il dataset originale viene diviso in modo casuale in due parti:

- **Training set** che contiene l'80% delle osservazioni;
- **Test set** che contiene il restante 20% del dataset originale.

Con i dati contenuti del traing set si effettua il fitting del modello logistico, mentre con il test set il modello appena creato viene validato. Al fine di testare la robustezza del modello, la creazione del training set e del test set è stata ripetuta più volte; i risultati riportati tengono conto della media di tutte le ripetizioni.
Infine, per valutare la bontà dei modelli proposti, le diagnosi stimate con i modelli di regressione logistica verranno confrontate con le vere diagnosi utilizzando diversi criteri:

- Soglia $p = 0.5$;
- Soglia $p = \bar{p}$;
- Tramite curva ROC.

Nei paragrafi successivi questi aspetti verrano descritti nel dettaglio, per ogni step dell'albero di classificazione.

4.1 Classificatore 1: Onda P

Per prima cosa sembra opportuno discriminare tra pazienti affetti da fibrillazione atriale e tutti gli altri. Per fare questo si imposta un modello di regressione logistica che ha come regressori l'età, il sesso e il flag binario che indica la presenza o meno di onda P. La scelta delle variabili da inserire nel modello è stata dettata dal risultato dell'algoritmo *stepwise*.
Viene creata una nuova variabile risposta $Y_i \sim Be(p_i)$ che divide il dataset in due popolazioni:

$$logit(p_i) = logit(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 \cdot Age_i + \beta_2 \cdot Sesso_i + \beta_3 \cdot OndaP_i \quad (4.1)$$

dove $Y_i = \begin{cases} 1 & : \text{ se il paziente } i \text{ è affetto da FA} \\ 0 & : \text{ altrimenti} \end{cases}$

Rispetto a quanto osservato nel Capitolo 2, ci si aspetta una forte significatività della variabile Onda.P.
Nel Listato 4.1 è riportato l'output del comando glm che effettua il fitting di modelli generalizzati in R ed in particolare la regressione logistica, specificando binomial come link function.

Listing 4.1: Stime dei p-values e dei parametri del modello (4.1)

```
Call:
  glm(formula = Diag ~ Age + Sesso + Onda.P_flag, family
      = "binomial",
      data = train)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.7110  -0.0523  -0.0398  -0.0272   4.0100

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept) 12.935081   1.251586  10.335  < 2e-16 ***
  Age         -0.027306   0.008613  -3.170 0.001523 **
  Sesso       -0.679349   0.186716  -3.638 0.000274 ***
  Onda.P_flag -8.606610   0.753886 -11.416  < 2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
      .1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2591.92  on 2454  degrees of freedom
Residual deviance:  757.25  on 2451  degrees of freedom
AIC: 765.25

Number of Fisher Scoring iterations: 9
```

Tutte le variabili sono molto significative, con p-value inferiori a 0.15 %, ma quella che influisce di più è sicuramente la presenza dell'onda P. Essa infatti risulta avere forte influenza negativa e poiché la presenza dell'onda P (cioè flag pari a 1) è prerogativa dei pazienti sani, il fatto che il parametro ad essa riferito sia negativo, giustifica la supposizione fatta all'inizio sul suo ruolo.

Tabelle di misclassificazione e curva ROC

Vengono riportate le Tabelle 4.1 e 4.2 di misclassificazione del modello (4.1).

	$\hat{0}$	$\hat{1}$
0	426	28
1	0	159

- Sensibilità = 100.00%
- Specificità = 93.83%
- Percentuale di classificazioni corrette = 95.43%

Tabella 4.1: Tabella di misclassificazione del modello (4.1) con soglia pari a 0.5

	$\hat{0}$	$\hat{1}$
0	426	28
1	0	159

- Sensibilità = 100.00%
- Specificità = 93.83%
- Percentuale di classificazioni corrette = 95.43%

Tabella 4.2: Tabella di misclassificazione del modello (4.1) con soglia pari a $\bar{p} = 0.28$

Si riporta in Figura 4.2 la curva ROC del primo classificatore costruita grazie al pacchetto ROCR, [14]. Dall'andamento del grafico si può affermare che risulta un ottimo classificatore. Infatti si nota che l'area sottesa dalla curva sia molto vicina a 1, e questo è indice di un classificatore altamente accurato, si veda la Tabella 3.2.

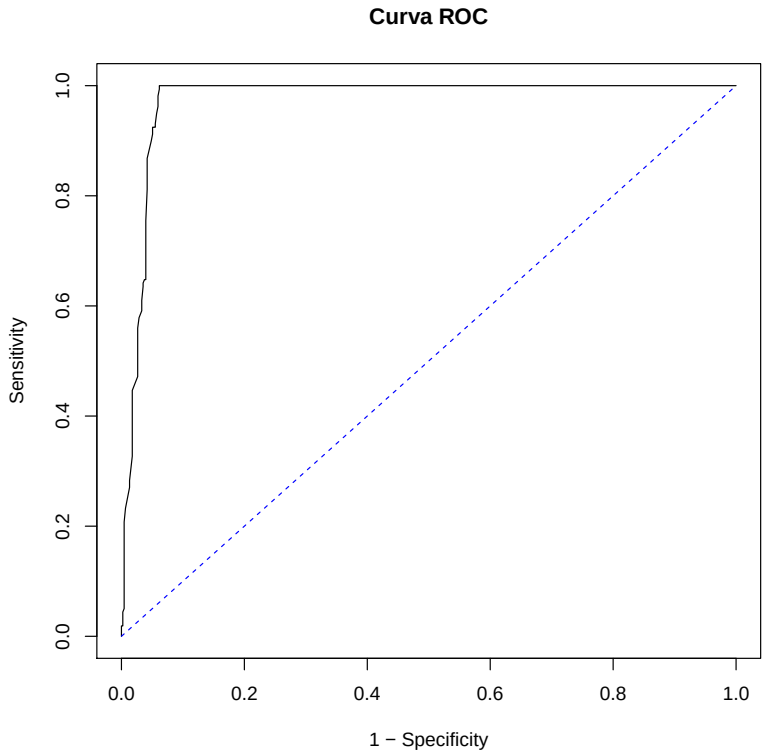


Figura 4.2: Curva ROC del modello (4.1)

4.2 Classificatore 2: Ampiezza QRS

A questo punto, si focalizza l'attenzione su un dataset ridotto, contenente solo pazienti RBBB, LBBB e sani. L'obiettvo ora è separare la classe dei sani da quella dei pazienti affetti da blocco di branca (sia destra che sinistra). Per fare questo si imposta un modello di regressione logistica che ha come regressori l'età, il sesso e la variabile che indica l'ampiezza del complesso QRS.
Viene creata una nuova variabile risposta $Y_i \sim Be(p_i)$ che divide il dataset

in due popolazioni:

$$logit(p_i) = logit(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 \cdot Age_i + \beta_2 \cdot Sesso_i + \beta_3 \cdot QRS_i \quad (4.2)$$

dove $Y_i = \begin{cases} 1 & : \text{ se il paziente } i \text{ è sano} \\ 0 & : \text{ se il paziente } i \text{ è affetto da BBB} \end{cases}$

Listing 4.2: Stime dei p-values e dei parametri del modello (4.2)

```
Call:
  glm(formula = Diag2 ~ Age + Sesso + QRS, family = "
      binomial",
       data = train1)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-3.7761  -0.0426   0.0963   0.2479   4.5154

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept) 23.167874   1.289998  17.960   <2e-16 ***
  Age          -0.086752   0.008245 -10.521   <2e-16 ***
  Sesso        -0.512486   0.235320  -2.178   0.0294 *
  QRS          -0.141153   0.008277 -17.054   <2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
      .1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2327.02  on 1893  degrees of freedom
Residual deviance:  573.91  on 1890  degrees of freedom
AIC: 581.91

Number of Fisher Scoring iterations: 7
```

Tutte le variabili sono molto significative, con p-values inferiori al 3 %. Fra tutte, quella che influisce di più è l'ampiezza del complesso QRS. Essa infatti risulta avere influenza negativa, come ci si aspettava. Infatti l'allargamento del QRS risulta essere un sintomo riscontrato nei pazienti affetti da blocco di branca, e poichè $Y_i = 0$ se il paziente i-esimo è malato, il coefficiente negativo è coerente con l'assunzione iniziale.

Tabelle di misclassificazione e curva ROC

Vengono riportate le Tabelle 4.3 e 4.4 di misclassificazione del modello (4.2).

	$\hat{0}$	$\hat{1}$
0	129	14
1	2	328

- Sensibilità = 99.39%
- Specificità = 90.21%
- Percentuale di classificazioni corrette = 96.62%

Tabella 4.3: Tabella di misclassificazione del modello (4.2) con soglia pari a 0.5

	$\hat{0}$	$\hat{1}$
0	133	13
1	2	325

- Sensibilità = 99.39%
- Specificità = 91.09%
- Percentuale di classificazioni corrette = 96.83%

Tabella 4.4: Tabella di misclassificazione del modello (4.2) con soglia pari a $\bar{p} = 0.67$

Si riporta in Figura 4.3 la curva ROC del secondo classificatore, con AUC=98.3%. Anche in questo caso il classificatore risulta molto accurato, come si può vedere anche dall'alta percentuale di classificazione corrette ottenute sul test-set, Tabella 4.3 e Tabella 4.4.

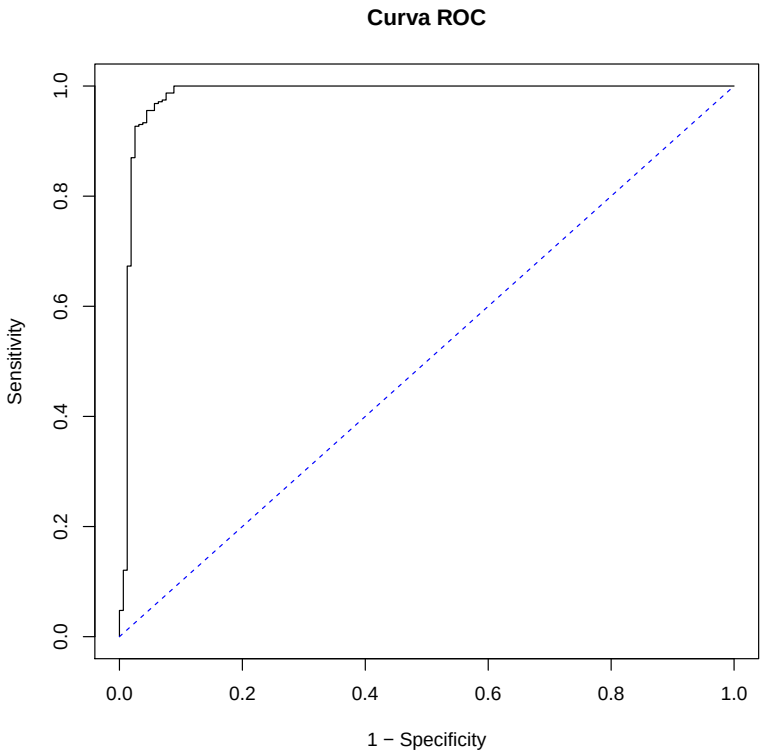


Figura 4.3: Curva ROC del modello (4.2)

4.3 Classificatore 3: Inversione onda T

Dopo aver isolato i soli pazienti con blocco di branca, resta da dividere all'interno di questo gruppo quali pazienti sono affetti da LBBB e quali da RBBB. Questo ultimo step è il più delicato, perché le differenze tra le due classi di patologie che emergono dalla Derivazione I dell'ECG sono lievi. La via scelta è quella di considerare l'inversione dell' onda T nella prima derivazione come elemento caratterizzante dei soggetti LBBB, come spiegato in [5].
Anche in questa situazione viene creata una nuova variabile risposta $Y_i \sim Be(p_i)$ che divide il dataset in due popolazioni:

$$logit(p_i) = logit(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 \cdot Age_i + \beta_2 \cdot Sesso_i + \beta_3 \cdot Inv.OndaT_i \quad (4.3)$$

dove $Y_i = \begin{cases} 1 & : \text{se il paziente i è afetto da RBBB} \\ 0 & : \text{se il paziente è affetto da LBBB} \end{cases}$

Listing 4.3: Stime dei p-values e dei parametri del modello (4.3)

```
Call:
  glm(formula = Diag4 ~ Age + Sesso + invA, family = "
     binomial",
      data = train2)

Deviance Residuals:
  Min       1Q   Median       3Q      Max
-1.9078  -0.8042   0.6230   0.7247   1.8860

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.719113   0.765095  -2.247  0.02464 *
  Age          -0.005102   0.008062  -0.633  0.52681
Sesso         0.600049   0.203352   2.951  0.00317 **
  invApiu       2.386638   0.203533  11.726  < 2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
     .1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 806.79  on 587  degrees of freedom
Residual deviance: 618.56  on 584  degrees of freedom
AIC: 626.56

Number of Fisher Scoring iterations: 4
```

Si nota subito che la variabile Age risulta non significativa. Questo è coerente con quanto affermato nel Capitolo 2, infatti le popolazioni sulle quali è stato creato questo modello sono solo quelle dei pazienti affetti da RBBB e LBBB, che come si può vedere nella Tabella 2.1 hanno un andamento molto simile della variabile Age. Anche in questo modello, come nei modelli (4.1) e (4.2) la variabile più significativa e con maggiore peso risulta essere quella che descrive una caratteristica peculiare della patologia, in questo caso l'inversione dell'onda T. La variabile, che vale 1 se non c'è inversione, ha coefficiente positivo, in accordo con quanto supposto.

Tabelle di misclassificazione e curva ROC

Vengono riportate le Tabelle 4.5 e 4.6 di misclassificazione del modello (4.3):

	$\hat{0}$	$\hat{1}$
0	38	13
1	9	86

- Sensibilità = 90.53%
- Specificità = 74.51%
- Percentuale di classificazioni corrette = 84.93%

Tabella 4.5: Tabella di misclassificazione del modello (4.3) con soglia pari a 0.5

	$\hat{0}$	$\hat{1}$
0	38	13
1	9	86

- Sensibilità = 90.53%
- Specificità = 74.51%
- Percentuale di classificazioni corrette = 84.93%

Tabella 4.6: Tabella di misclassificazione del modello (4.3) con soglia pari a $\bar{p} = 0.45$

Si riporta in Figura 4.4 la curva ROC del terzo classificatore, che risulta meno accurato dei due precedenti, con un AUC=84.1%.

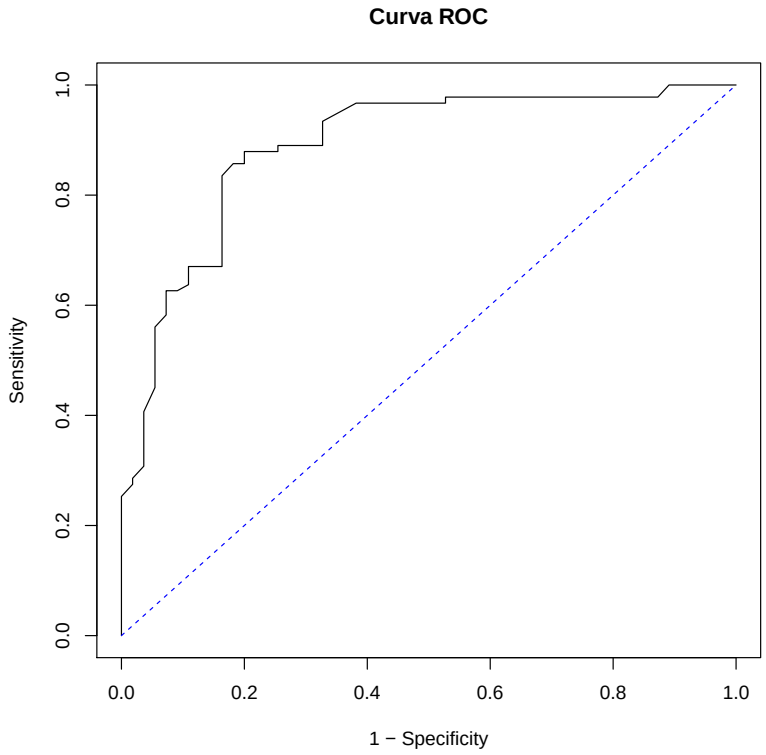


Figura 4.4: Curva ROC del modello (4.3)

I risultati ottenuti mostrano che la presenza o meno dell'onda P e la variazione in ampiezza del complesso QRS, sono variabili che permettono di costruire buoni classificatori (con percentuale di classificazioni corrette mediamente intorno al 95 %) rispettivamente per discriminare tra soggetti affetti da fibrillazione atriale e tutti gli altri e per discriminare fra sani e pazienti con blocco di branca.
Il terzo classificatore, che ha l'obiettivo di distinguere fra pazienti LBBB e RBBB, risulta mediamente accurato, con perentuale di soggetti classificati correttamente pari all' 85 %.

4.4 Dati normalizzati

Analogamente a quanto fatto nelle Sezioni 4.1, 4.2, 4.3, sono riportate in questa parte le tabelle di misclassificazione e i grafici delle curve ROC, applicando i modelli (4.1), (4.2), (4.3) ad un dataset normalizzato.

Per dataset normalizzato si intende il detaset in cui le ampiezze dei 3 segmenti che costituiscono tracciato elettrocardiografico (onda P, complesso QRS e onda T) sommino a 1. In questo modo la variabile non assume più il significato di "ampiezza", ma di peso percentuale sul totale del tracciato. Questo nuovo dataset è stato creato con lo scopo di ricondurre tutti i dati su una scala temporale di riferimento, per eliminare eventuali problemi di variabilità di fase.

4.4.1 Classificatore 1: Onda P

Vengono riportate le Tabelle 4.7 e 4.8 di misclassificazione del modello (4.1):

	$\hat{0}$	$\hat{1}$
0	441	30
1	2	140

- Sensibilità = 98.59%
- Specificità = 93.63%
- Percentuale di classificazioni corrette = 94.78%

Tabella 4.7: Tabella di misclassificazione del modello (4.1) con dati normalizzati e soglia pari a 0.5

	$\hat{0}$	$\hat{1}$
0	441	30
1	2	140

- Sensibilità = 98.59%
- Specificità = 93.63%
- Percentuale di classificazioni corrette = 94.78%

Tabella 4.8: Tabella di misclassificazione del modello (4.1) con dati normalizzati e soglia pari a $\bar{p}$

Si riporta in Figura 4.5 la curva ROC del primo classificatore.

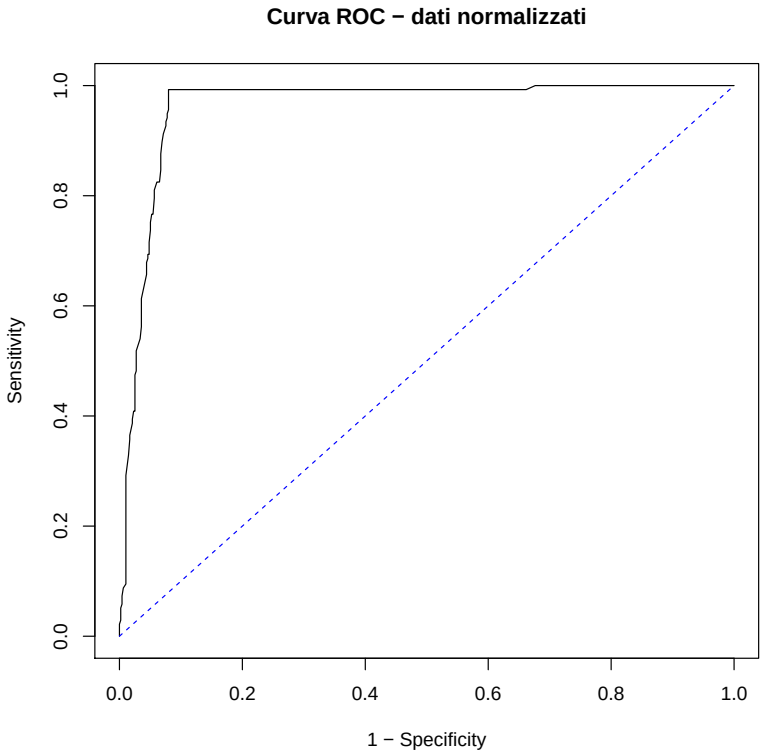


Figura 4.5: Curva ROC del modello (4.1) con dati normalizzati

4.4.2 Classificatore 2: Ampiezza QRS

Vengono riportate le Tabelle 4.9 e 4.10 di misclassificazione del modello (4.2):

	$\hat{0}$	$\hat{1}$
0	134	16
1	2	321

- Sensibilità = 93.38%
- Specificità = 89.34%
- Percentuale di classificazioni corrette = 96.19%

Tabella 4.9: Tabella di misclassificazione del modello (4.2) con dati normalizzati e soglia pari a 0.5

	$\hat{0}$	$\hat{1}$
0	139	11
1	12	311

- Sensibilità = 96.28%
- Specificità = 92.66%
- Percentuale di classificazioni corrette = 95.14%

Tabella 4.10: Tabella di misclassificazione del modello (4.2) con dati normalizzati e soglia pari a $\bar{p}$

Si riporta in Figura 4.6 la curva ROC del secondo classificatore.

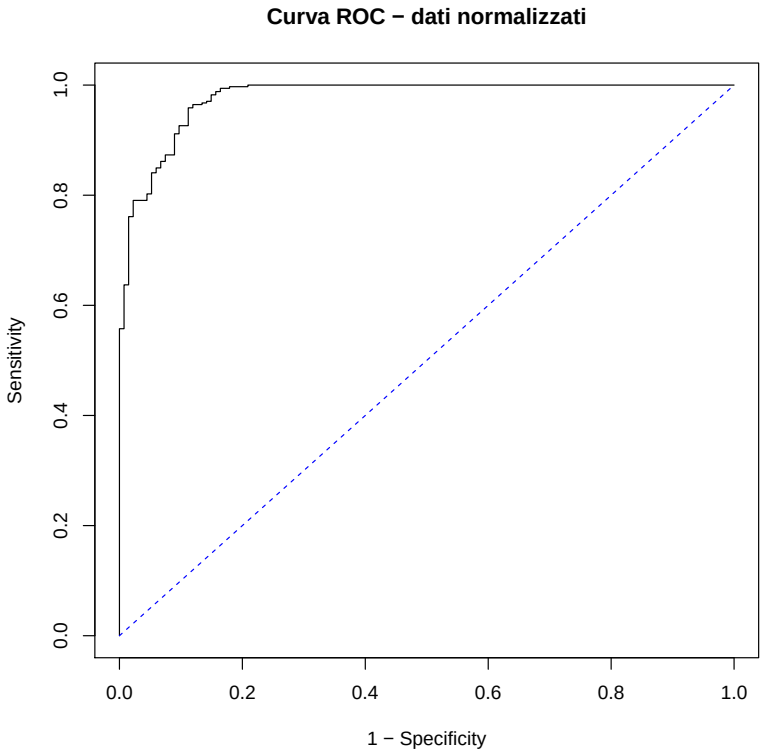


Figura 4.6: Curva ROC del modello (4.2) con dati normalizzati

4.4.3 Classificatore 3: Inversione onda T

Vengono riportate le Tabelle 4.11 e 4.12 di misclassificazione del modello (4.3):

	$\hat{0}$	$\hat{1}$
0	50	12
1	15	69

- Sensibilità = 82.14%
- Specificità = 80.65%
- Percentuale di classificazioni corrette = 81.51%

Tabella 4.11: Tabella di misclassificazione del modello (4.3) con dati normalizzati e soglia pari a 0.5

	$\hat{0}$	$\hat{1}$
0	50	12
1	15	69

- Sensibilità = 82.14%
- Specificità = 80.65%
- Percentuale di classificazioni corrette = 81.51%

Tabella 4.12: Tabella di misclassificazione del modello (4.3) con dati normalizzati e soglia pari a $\bar{p}$

Si riporta in Figura 4.7 la curva ROC del terzo classificatore.

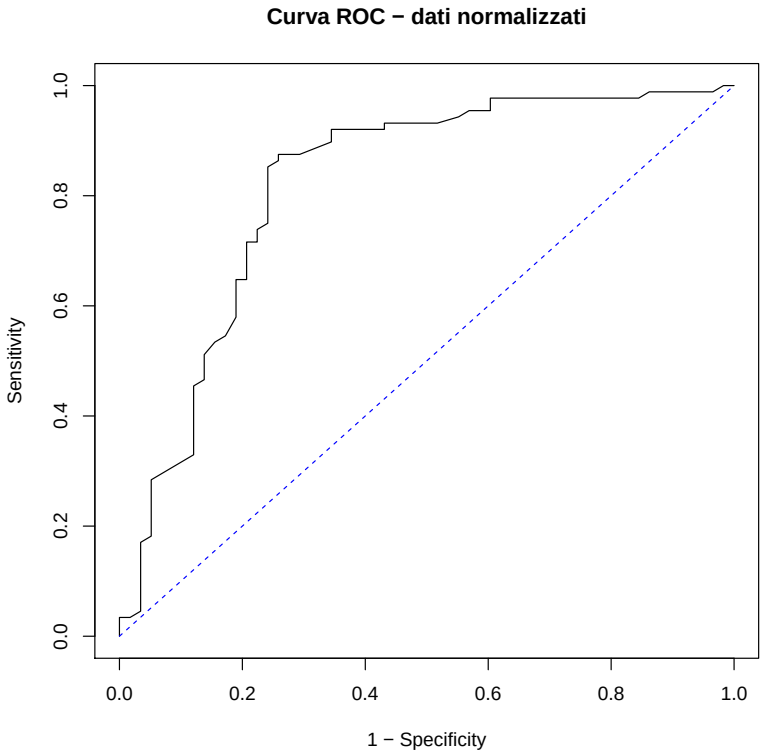


Figura 4.7: Curva ROC del modello (4.3) con dati normalizzati

Si nota che i risultati ottenuti non sono uniformemente migliori di quelli registrati nel caso di applicazione degli stessi modelli al dataset non standardizzato. Per questo motivo è stato deciso di utilizare il dataset orignale, di più facile interpretazione, nello sviluppo dell'applicazione descritta nel Capitolo 5.

Capitolo 5

Sviluppo di un'applicazione GUI di classificazione automatica

In questo capitolo viene presentata un'applicazione web **MyDoctor**, disponibile all'indirizzo https://silvia.shinyapps.io/MyDoctor/, creata per permettere all'utente finale di utilizzare i metodi proposti in modo facile e veloce, senza doversi preoccupare dei dettagli implementativi e senza dover disporre di particolari software.
Questo è stato possibile grazie al pacchetto "Shiny" di RStudio, [1], che permette di creare interfacce grafiche interattive, scrivendo solo codice R. Grazie all'introduzione di widgets pre-costruiti è possibile interagire direttamente con l'ambiente R, ad esempio caricando il database sul quale verrà poi effettuata l'analisi, oppure scegliendo dei parametri da un elenco a scelta multipla, che diventeranno gli input dell'algoritmo. Per ulteriori informazioni si veda [16].
Uno dei principali requisiti della statistica è quello di **comunicare**, cioè di riuscire a processare grandi moli di dati, analizzarli, interpretarli e da questi estarre le informazioni di interesse per coloro che hanno commissionato l'analisi. In questo caso si suppone che l'utilizzatore finale dei metodi descritti in questa tesi sia il personale medico; per questo motivo è necessario riuscire a tradurre il "linguaggio statstico"in un linguaggio più accessibile.
Shiny si colloca perfettamente in questo contesto: permette infatti di rendere accattivante, intuitivo e condivisibile un algoritmo, senza dover rinunciare alle potenzialità del calcolo di R.

5.1 MyDoctor

Per costruire un'applicazione con Shiny è necessario installare RStudio e il pacchetto Shinyapps, [17]. Dal punto di vista implementativo l'applicazione è costituita da due file che devono essere contenuti nella stessa cartella:

- ui.R;
- server.R.

5.1.1 File `ui.R`

Il File `ui.R` è la descrizione dell'interfaccia garfica: viene definito il layout della pagina, vengono inseriti i widget e i comandi di visualizzazione dell'output. Attraverso i widgets (checkbox input, file upload control, select list input control, ecc...) vengono richiesti all'utente i valori in ingresso, oppure la selezione del tipo di output desiderato. L'output può essere costituito da molteplici elementi, tra cui grafici, tabelle, immagini e caselle di testo. Si riporta nel Listato 5.1 il codice del File ui.R per l'applicazione **MyDoctor**.

Listing 5.1: File `ui.R`

```
shinyUI(pageWithSidebar(
  headerPanel("MyDoctor"),
  sidebarPanel(
    fileInput('file1', 'Choose ECG File',
              accept=c('text/csv', 'text/comma-separated
                 -values,text/plain', '.txt')),
    tags$hr(),
    checkboxInput('header', 'Header', TRUE),
    radioButtons('sep', 'Separator',
                 c(Comma=',',
                   Semicolon=';',
                   Tab='	'),
                 ','),
    radioButtons('quote', 'Quote',
                 c(None='',
                   'Double Quote'='"',
                   'Single Quote'="'"),
                 '"')
  ),
  mainPanel(
    tableOutput('contents'),
    br(),
    plotOutput("tracciato"),
    br(),
    h4(textOutput("text1")),
    br(),
```

```
26      h4(textOutput("text2")),
27      br(),
28      h4(textOutput("text3")),
29      br(),
30      h4(textOutput("text4"))
31    )
32  ))
```

5.1.2 File `server.R`

Il secondo file necessario per il funzionamento di un'applicazione creata con Shiny, è `server.R`, che contiene la parte dell'applicazione che "manipola"i dati. Nel complesso, in questo file sono contenuti due tipi di dati:

- **Reactive objects:** dataset che vengono importati dall'input definto dall'utente e poi trasferiti alle istruzioni che gestiscono i diversi output;
- **Output:** il comando "Output" introduce il codice che gestisce l'elaborazione dei dati e la produzione del risultato desiderato, sia esso un grafico, una tabella o una stringa. La variabile *output* viene poi richiamata nel file `ui.R` in modo che possa essere visualizzata nell'interfaccia grafica.

Si riporta nel Listato 5.1 il codice del File server.R per l'applicazione **MyDoctor**.

Listing 5.2: File `server.R`

```
1  shinyServer(function(input, output) {
2
3   output$contents <- renderTable({
4   inFile <- input$file1
5   if (is.null(inFile))
6   return(NULL)
7   read.csv(inFile$datapath, header=input$header, sep=input
        $sep, quote=input$quote)
8  })
9
10 output$text1 <- renderText({
11   inFile <- input$file1
12   if (is.null(inFile))
13   return(NULL)
14   data <- read.table(inFile$datapath, header=input$header
         , sep=input$sep, quote=input$quote)
15   n=dim(data)[1]
16   p=dim(data)[2]
17   c1<- read.table("coefR1.txt",header=TRUE)
18   c2<- read.table("coefR2.txt",header=TRUE)
19   c3<- read.table("coefR3.txt",header=TRUE)
```

```
20    prob1=exp(c1[,1][1]+ c1[,1][2]*as.numeric(data$Age) +c1
          [,1][3]*as.numeric(data$Sesso) + c1[,1][4]*
          as.numeric(data$Onda.P_flag))/(1 + exp(c1[,1][1]+
          c1[,1][2]*as.numeric(data$Age) +c1[,1][3]*
          as.numeric(data$Sesso) + c1[,1][4]*as.numeric(data$
          Onda.P_flag)))
21    prob2=exp(c2[,1][1]+ c2[,1][2]*as.numeric(data$Age) +c2
          [,1][3]*as.numeric(data$Sesso) + c2[,1][4]*
          as.numeric(data$QRS))/(1 + exp(c2[,1][1]+ c2[,1][2]
          *as.numeric(data$Age) +c2[,1][3]*as.numeric(data$
          Sesso) + c2[,1][4]*as.numeric(data$QRS)))
22    prob3=exp(c3[,1][1]+ c3[,1][2]*as.numeric(data$Age) +c3
          [,1][3]*as.numeric(data$Sesso) + c3[,1][4]*
          as.numeric(data$invA))/(1 + exp(c3[,1][1]+ c3
          [,1][2]*as.numeric(data$Age) +c3[,1][3]*as.numeric(
          data$Sesso) + c3[,1][4]*as.numeric(data$invA)))
23    if(prob1>=0.5) diagnosi= "Fibrillazione atriale"
24    if(prob1<0.5 & prob2>=0.5) diagnosi="Sano"
25    if(prob1<0.5 & prob2<0.5 & prob3>=0.5) diagnosi="RBBB"
26    if(prob1<0.5 & prob2<0.5 & prob3<0.5) diagnosi="LBBB"
27    paste("Il paziente selezionato risulta essere:",
          diagnosi)
28  })
29
30  output$text2 <- renderText({
31    inFile <- input$file1
32    if (is.null(inFile))
33    return(NULL)
34    data <- read.table(inFile$datapath, header=input$header
          , sep=input$sep, quote=input$quote)
35    c1<- read.table("coefR1.txt",header=TRUE)
36    prob1=exp(c1[,1][1]+ c1[,1][2]*as.numeric(data$Age) +c1
          [,1][3]*as.numeric(data$Sesso) + c1[,1][4]*
          as.numeric(data$Onda.P_flag))/(1 + exp(c1[,1][1]+
          c1[,1][2]*as.numeric(data$Age) +c1[,1][3]*
          as.numeric(data$Sesso) + c1[,1][4]*as.numeric(data$
          Onda.P_flag)))
37    paste("Probabilita' di essere affetto da Fibrillazione
          Atriale=", signif(prob1,3))
38  })
39
40  output$text3 <- renderText({
41    inFile <- input$file1
42    if (is.null(inFile))
43    return(NULL)
44    data <- read.table(inFile$datapath, header=input$header
          , sep=input$sep, quote=input$quote)
45    c2<- read.table("coefR2.txt",header=TRUE)
```

```
46    prob2=exp(c2[,1][1]+ c2[,1][2]*as.numeric(data$Age) +c2
         [,1][3]*as.numeric(data$Sesso) + c2[,1][4]*
         as.numeric(data$QRS))/(1 + exp(c2[,1][1]+ c2[,1][2]
         *as.numeric(data$Age) +c2[,1][3]*as.numeric(data$
         Sesso) + c2[,1][4]*as.numeric(data$QRS)))
47    c1<- read.table("coefR1.txt",header=TRUE)
48    prob1=exp(c1[,1][1]+ c1[,1][2]*as.numeric(data$Age) +c1
         [,1][3]*as.numeric(data$Sesso) + c1[,1][4]*
         as.numeric(data$Onda.P_flag))/(1 + exp(c1[,1][1]+
         c1[,1][2]*as.numeric(data$Age) +c1[,1][3]*
         as.numeric(data$Sesso) + c1[,1][4]*as.numeric(data$
         Onda.P_flag)))
49    if(prob1<0.5)  paste("Probabilita' di essere sano=",
         signif(prob2,3))
50    else return(NULL)
51  })
52
53  output$text4 <- renderText({
54    inFile <- input$file1
55    if (is.null(inFile))
56    return(NULL)
57    data <- read.table(inFile$datapath, header=input$header
         , sep=input$sep, quote=input$quote)
58    c3<- read.table("coefR3.txt",header=TRUE)
59    prob3=exp(c3[,1][1]+ c3[,1][2]*as.numeric(data$Age) +c3
         [,1][3]*as.numeric(data$Sesso) + c3[,1][4]*
         as.numeric(data$invA))/(1 + exp(c3[,1][1]+ c3
         [,1][2]*as.numeric(data$Age) +c3[,1][3]*as.numeric(
         data$Sesso) + c3[,1][4]*as.numeric(data$invA)))
60    c2<- read.table("coefR2.txt",header=TRUE)
61    prob2=exp(c2[,1][1]+ c2[,1][2]*as.numeric(data$Age) +c2
         [,1][3]*as.numeric(data$Sesso) + c2[,1][4]*
         as.numeric(data$QRS))/(1 + exp(c2[,1][1]+ c2[,1][2]
         *as.numeric(data$Age) +c2[,1][3]*as.numeric(data$
         Sesso) + c2[,1][4]*as.numeric(data$QRS)))
62    c1<- read.table("coefR1.txt",header=TRUE)
63    prob1=exp(c1[,1][1]+ c1[,1][2]*as.numeric(data$Age) +c1
         [,1][3]*as.numeric(data$Sesso) + c1[,1][4]*
         as.numeric(data$Onda.P_flag))/(1 + exp(c1[,1][1]+
         c1[,1][2]*as.numeric(data$Age) +c1[,1][3]*
         as.numeric(data$Sesso) + c1[,1][4]*as.numeric(data$
         Onda.P_flag)))
64    if(prob1<0.5 & prob2<0.5)  paste("Probabilita' di
         essere affetto da RBBB=", signif(prob3,3))
65    else return(NULL)
66    })
67
68  output$tracciato <- renderPlot({
69    inFile <- input$file1
```

```
70    if (is.null(inFile))
71    return(NULL)
72    data <- read.table(inFile$datapath, header=input$header
          , sep=input$sep, quote=input$quote)
73    datafun=data[12:1211]
74    x <- seq(from=1, to=1200, by=1)
75    y <- datafun
76    plot(x, y, col = 'red', type='l', main="Derivazione I",
          xlab="Tempo [ms]", ylab="Tensione [mV]")
77  })
78
79  })
```

Si riportano in Figura 5.1, Figura 5.2 e Figura 5.3 alcune schermate dei momenti salienti dell'uso di **MyDoctor**.

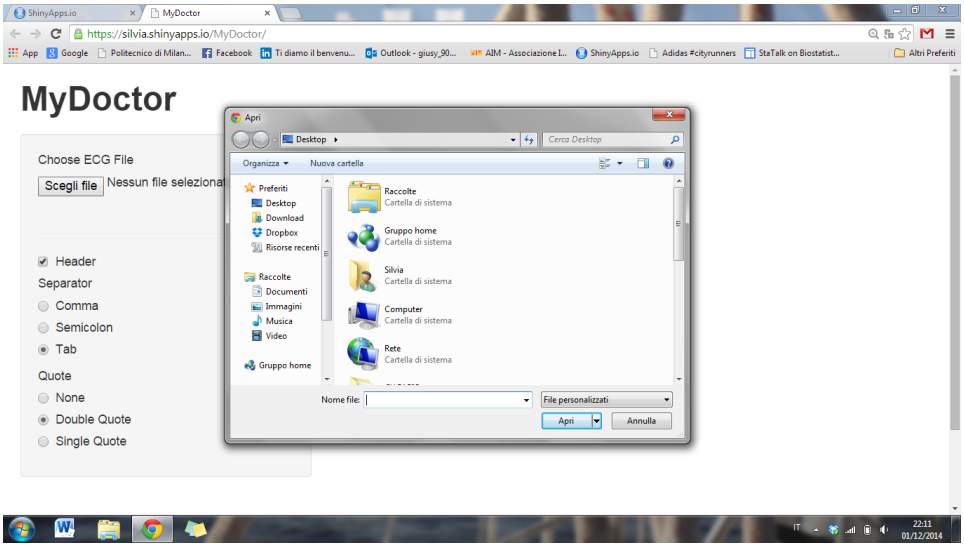


Figura 5.1: Interfaccia grafica 1/3

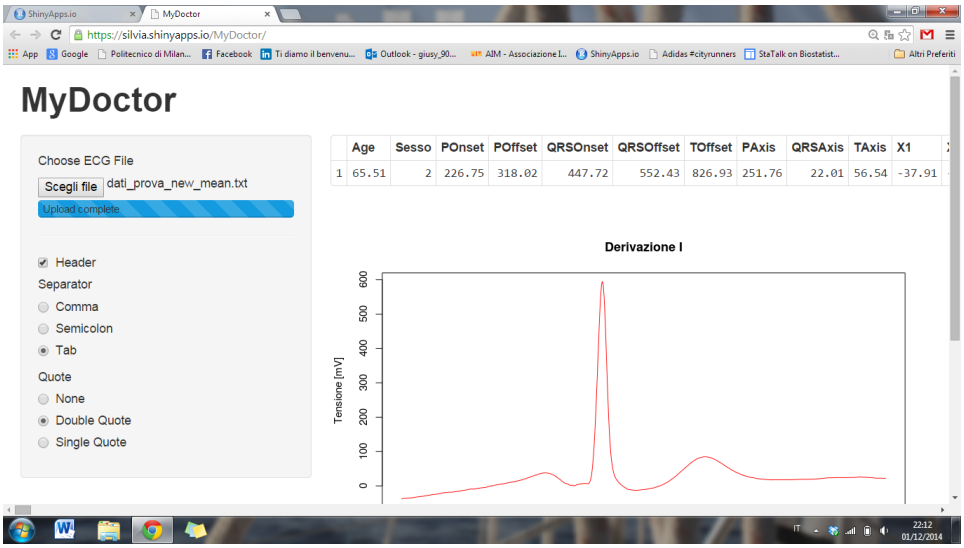


Figura 5.2: Interfaccia grafica 2/3

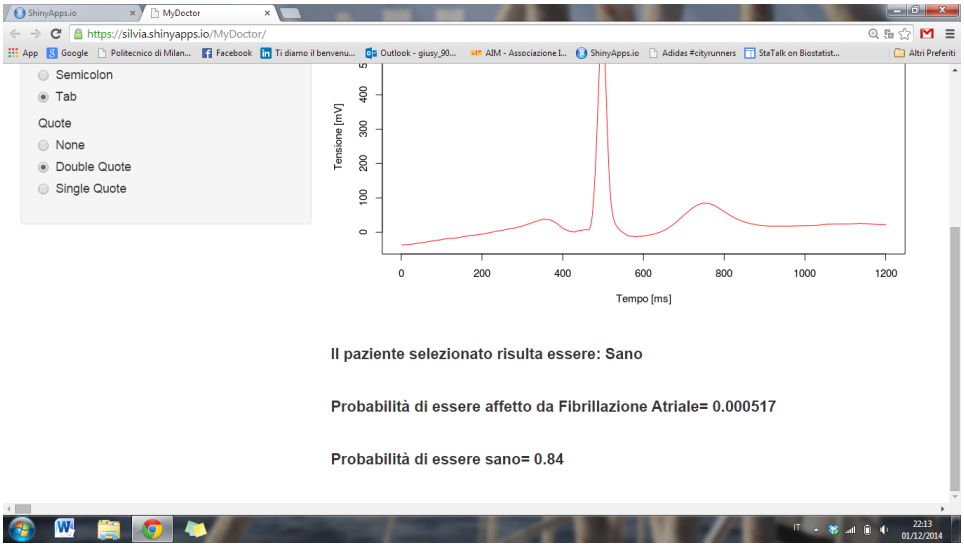


Figura 5.3: Interfaccia grafica 3/3

Capitolo 6

Conclusioni e sviluppi futuri

In questa tesi ci si è concentrati su una proposta di classifficazione supervisionata di dati multivariati, in cui ogni osservazione coincide con una serie di landmarks e variabli descrittive dello stato di salute di ogni paziente. Il dataset a disposizione consiste in un insieme di tracciati elettrocardiografici, fisiologici e patologici, per i quali sono registrate otto delle dodici derivazioni standard. Ci si è concentrati su quattro classi di pazienti: sani affetti da blocco di branca destra, blocco di branca sinistra e da fibrillazione atriale. Uno dei principali obiettivi è stato quello di prevedere l'eventuale presenza di una patologia nei soggetti e per far questo è stato creato un albero decisionale, dove ad ogni step è stata utilizzata la regressione logistica per stimare le probabilità di appartenenza di ogni paziente ad una data classe. I risultati ottenuti analizzando le features della derivazione I mostrano che la presenza o meno dell'onda P e la variazione in ampiezza del complesso QRS, sono variabili che permettono di costruire buoni classificatori (con percentuale di classificazioni corrette mediamente intorno al 95 %) rispettivamente per discriminare tra soggetti affetti da fibrillazione atriale e tutti gli altri e per discriminare fra sani e pazienti con blocco di branca.
La fase più delicata della classificazione è stata quella riguardante la distinzione fra pazienti LBBB e RBBB, dove la perentuale di soggetti classificati correttamente scende all' 85 %. Le cause sono molteplici: in primo luogo il dateset con il quale sono state condotte le analisi è fortemente sbilanciato (con 1,633 soggetti sani e solo 420 RBBB e 314 LBBB). Questo comporta una difficoltà maggiore nella definizione del training e del test set, dovuta alla presenza di un trade-off tra numerosità del campione su cui fare test e massimizzazione della capacità di distinguere del modello stesso. Il secondo problema può essere ricondotto all'assenza di smoothing del dato funzionale; la variabile inserita nel modello che discrimina tra RBBB e LBBB è il "flag"che assume valore 1 se c'è inversione dell'onda T e 0 altrimenti. L'inversione dell'onda T dipende dall'area sottesa dalla curva stessa, e per questa ragione l'utilizzo di dati precedentemente sottoposti ad una fase di

filtraggio e smoothing, renderebbe l'informazione sull'area più precisa e attendibile.
A questo proposito, uno dei prossimi obiettivi sarà quello di studiare approfonditamente le altre derivazioni dell'ECG al fine di ricercare pattern specifici, rappresentativi di una determinata patologia, che possano essere utilizzati per migliorare la classificazione. Nel dettaglio, si dovrebbe porre particolare attenzione alla derivazione V1 e alla derivazione V6, per migliorare la capacità di discriminazione tra RBBB e LBBB, che allo stato attuale risulta essere la meno efficace. Come descritto in letteratura medica (si veda [18]), esistono infatti alcune caratteristiche peculiari dei soggetti affetti da blocco di branca, come ad esempio la comparsa di una seconda R, detta seconda onda R, che si presenta nelle derivazioni V1 e V2 per i pazienti RBBB e nelle derivazione V5 e V6 per i pazienti LBBB. Con questo ulteriore strumento si auspica di poter giungere ad una classificazione con errore di misclassficazione minimo all'interno della classe di patologie del blocco di branca.

Bibliografia

[1] RStudio e Inc. *shiny: Web Application Framework for R*. R package version 0.10.2.1. 2014. URL: http://CRAN.R-project.org/package=shiny.

[2] Francesca Ieva et al. ≪Multivariate functional clustering for the morphological analysis of electrocardiograph curves≫. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62.3 (2013), pp. 401–418.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2014. URL: http://www.R-project.org/.

[4] Dee Unglaub Silverthorn et al. *Human physiology: an integrated approach*. Pearson/Benjamin Cummings, 2009.

[5] Fabrizio Castaldo. *ECG Facile: Guida all'Interpretazione dell'ECG*. Antonio Delfino Editore, 2014.

[6] Umberto Gnudi. *Corso di elettrocardiografia di base*. 2012.

[7] E Piccolo et al. *L'onda T invertita, normale e minacciosa*. 2014.

[8] Myles Hollander, Douglas A Wolfe e Eric Chicken. *Nonparametric statistical methods*. Vol. 751. John Wiley & Sons, 2013.

[9] Howard Levene. ≪Robust tests for equality of variances1≫. In: *Contributions to probability and statistics: Essays in honor of Harold Hotelling* 2 (1960), pp. 278–292.

[10] NIST/SEMATECH e-Handbook of Statistical Methods. *Levene Test for Equality of Variances*. http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htmSect.1.3.5.10. 2014.

[11] Joseph L. Gastwirth et al. *lawstat: An R package for biostatistics, public policy, and law*. R package version 2.4.1. 2013. URL: http://CRAN.R-project.org/package=lawstat.

[12] Richard Arnold Johnson e Dean W Wichern. *Applied multivariate statistical analysis*. Vol. 4. Pearson Education, 1992.

[13] John A Swets. ≪Measuring the accuracy of diagnostic systems≫. In: *Science* 240.4857 (1988), pp. 1285–1293.

[14] T. Sing et al. ≪ROCR: visualizing classifier performance in R≫. In: *Bioinformatics* 21.20 (2005), p. 7881. URL: http://rocr.bioinf.mpi-sb.mpg.de.

[15] John R Hampton. *The ECG made easy*. Elsevier Health Sciences, 2013. Cap. 4.

[16] Chris Beeley. *Web Application Development with R Using Shiny*. Packt Publishing Ltd, 2013.

[17] JJ Allaire. *shinyapps: Interface to ShinyApps*. R package version 0.3.62. 2013.

[18] Letizia Vola e Paola Turina. *Guida pratica per l'interpretazione e la lettura di un elettrocardiogramma*. 2013.