Arrhythmia UCI Dataset
Report per l'Esame di Fondamenti di Machine Learning

Nicola Gutierrez

146875
Ingegneria informatica - sede di Mantova
276269@studenti.unimore.it

Abstract

L'obiettivo è quello di creare un modello di classificazione che tragga le conclusioni dai dati forniti distinguendo tra l'assenza o la presenza di aritmia cardiaca e di, eventualmente, classificarla in una delle 16 classi.

1 Introduzione

I dati utilizzati nel progetto sono stati donati da H. Altay Guvenir e il dataset è disponibile nell'UCI Machine Learning Repository sotto il nome di Arrhythmia Data Set. Il dataset presenta **452** samples, ognuno dei quali rappresenta la cartella di un paziente. Per quanto riguarda le features sono **279**, di cui 206 hanno valore lineare e 73 hanno valore nominale.
La feature ***class*** è la feature target e si suddivide come segue :

Nr	Nome della classe
1	Normal (absence of arrhythmia)
2	Ischemic changes (CAD)
3	Old Anterior Myocardial Infraction
4	Old Inferior Myocardial Infraction
5	Sinus tachycardy
6	Sinus bradycardy
7	Ventricular Premature Contraction (PVC)
8	Supraventricular Premature Contraction
9	Left Boundle branch block
10	Right boundle branch block
11	1.Degree AtrioVentricular block
12	2.Degree AV block
13	3.Degree AV block
14	Left Ventricule hypertrophy
15	Atrial Fibrillation or Flutter
16	Other

Si tratta di un caso di classificazione multiclass. Le classi *1.Degree AtrioVentricular block*, *2.Degree AV block* e *3.Degree AV block* non vengono rappresentate da nessun paziente presente all'interno del dataset, di conseguenza verranno implicitamente ignorate. La classe *Other* viene considerata nonostante la sua natura indefinita in modo da non ridurre ulteriormente il numero di samples; verrà considerata secondo il *worst-case concept* quindi verrà fatta ricadere sotto la macro-classe aritmia.

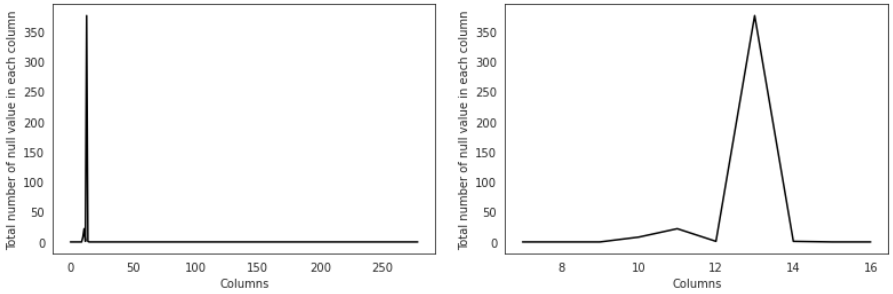
1.1 Gestione del dataset

Il dataset viene trattato in fase di Exploratory Data Analysis & Pre-processing nella sua interezza. Solo in seguito viene suddiviso in due sezioni dove la prima rappresenta l'**80%** del dataset ed è dedicata alla fase di training, mentre la seconda rappresenta il rimanente **20%** ed è dedicata alla fase di testing. La suddivizione avviene in seguito all'applicazione sui dati di un *random_state* (con seme 42) e attraverso l'opzione *stratify* in modo da assicurarsi una distribuzione maggiormente bilanciata delle classi tra i subset.

2 Exploratory Data Analysis & Pre-Processing

2.1 Gestione valori mancanti

Secondo le informazioni fornite insieme al dataset, i valori mancanti sono rappresentati con il simbolo '?'.



(a) visione di insieme. (b) vista dell'intorno del picco.

Figure 1: valori mancanti per ogni colonna.

Le classi che presentano valori mancanti sono *T*, *P*, *J*, *QRST* e *Heart Rate*.

Nome	**Valori mancanti**	**Soluzione**
T	8	Sostituzione attraverso tecniche di *imputation*
P	22	Sostituzione attraverso tecniche di *imputation*
J	1	Sostituzione attraverso tecniche di *imputation*
QRST	376	Feature scartata, l'**83%** dei suoi valori risulta mancante
Heart Rate	1	Sostituzione attraverso tecniche di *imputation*

In particolare, si è scelta la modalità più immediata per sostituire i valori indefiniti, cioè con la mediana dei valori della colonna, data la bassa percentuale di valori da determinare.

2.2 Gestione outliers

Dal controllo sulla presenza di possibili valori non veritieri si è riscontrato che nella maggior parte del dataset i valori sono alquanto uniformi e plausibili, l'unico caso degno di nota è stato riportato di seguente, dove si può notare la presenza di alcuni valori fuori scala all'interno degli attributi *Height* e *Weight*. Gli ulteriori grafici utilizzati sono all'interno della cartella *extra/outliers*.

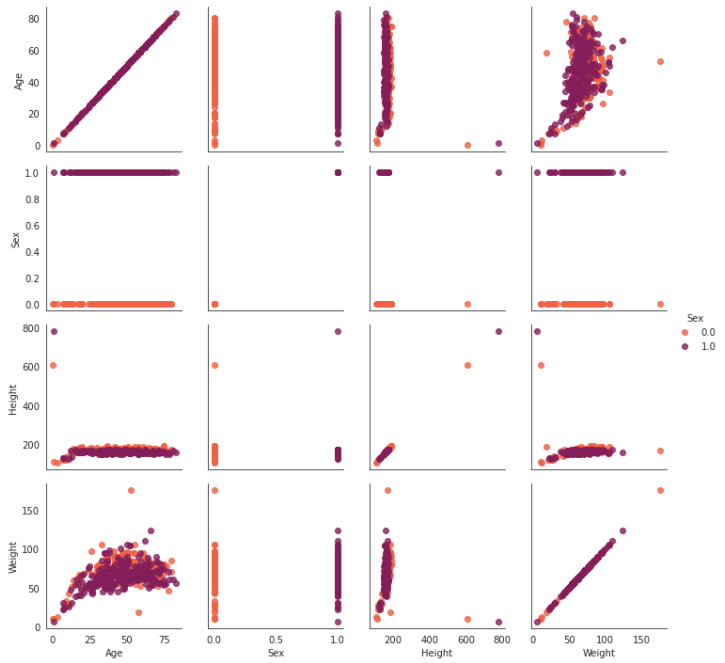


Figure 2: Controllo presenza di outliers.

2.3 Analisi distribuzione e correlazione feature

Analizzando la distribuzione totale delle feature ***class*** si può notare come il dataset sia **sbilanciato** verso i casi *normali* e come le classi riferite all'*aritmia* non siano particolarmente equilibrate. In generale, la distribuzione binaria tra casi *normali* e l'insieme dei casi di *aritmia* è piuttosto equilibrata (plot in *extra/class_distribution*). Con queste premesse una soluzione sarebbe quella di ridimensionare il dataset, eliminando o aggiungendo samples in modo da avere una distribuzione equa. Tuttavia, la prima opzione (*eliminare*) non è stata presa in considerazione in quanto il dataset offre di per sé pochi samples (se considerati rispetto alle feature), mentre la seconda (*aggiungere*) semplicemente non è possibile.

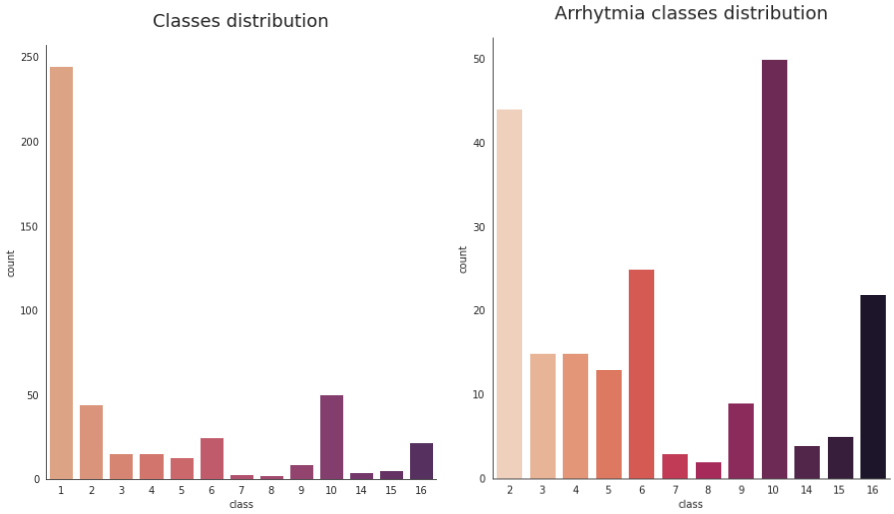


Figure 3: Distribuzione delle classi

Per quanto riguarda la correlazione tra feature :

- **185** coppie di attributi hanno una correlazione maggiore di $|0.7|$;
- **76** coppie di attributi hanno una correlazione maggiore di $|0.8|$;
- **19** coppie di attributi hanno una correlazione maggiore di $|0.9|$;

Sulla base di questi risultati si è deciso di rimuovere le **19** coppie con correlazione maggiore. Tra queste coppie solo la feature *AVF212* si ripropone due volte, quindi in totale vengono rimosse **18** feature, portando il numero di attributi del nuovo dataset a **261**

3 Selezione del modello

3.1 Modelli candidati

- **K-Nearest Neighbours** : scelto in quanto semplice da implementare e veloce nell'esecuzione dato il numero accessibile di samples;
- **Decision Tree Classifier** : scelto perché implicitamente potrebbe ridurre il numero di feature (*feature selection implicita*) e ottenere una buona *accuracy*, l'implementazione richiede poche risorse e risulta veloce a *inference time*;
- **Support Vector Machine** : scelto come riferimento in caso in cui si presentasse overfitting. Inoltre si prevede di utilizzare la libreria `sklearn`, semplificando la selezione degli iperparametri;
- **Softmax Regression** : scelto per completezza e per permettere un confronto completo anche in ottica di *ensemble*;

3.2 Cross validation

Per scegliere la migliore combinazione di iperparametri per ogni modello si è scelto l'approccio *k-fold*. In particolare, è stata utilizzata la funzione `GridSearchCV` di `sklearn` con $cv = 7$ e $score = accuracy$. Le configurazioni di iperparametri sono state scelte euristicamente per ogni modello. I risultati della *Grid Search* sono i seguenti :

- **K-Nearest Neighbours** : `n_neighbors` : **3**
- **Decision Tree Classifier** : `max_depth` : **10**, `criterion` : **gini**
- **Support Vector Machine** : `C` : **100**, `gamma` : **0.001**, `kernel` : **rbf**
- **Softmax Regression** : `penalty` : **l1**, `C` : **1.0**

3.3 Ensemble : Stacking Classifier

Oltre ai modelli sopra citati si è deciso di utilizzare il meta-algoritmo *stacking*, combinando i tre modelli con l'*accuracy* maggiore, quindi *SVM*, *K-NN* e *Decision Tree Classifier*.

3.4 Valutazione dei modelli

La valutazione è avvenuta tramite l'utilizzo della funzione `cross_validate` di `sklearn`.
In questa fase per scegliere il modello migliore si è preso in considerazione anche l'**F1-Score**, in quanto permette di effettuare valutazioni considerando più metriche (*Recall* e *Precision*).
Il modello finale scelto secondo quanto detto è lo ***Stacking Model***.

Modello	Accuracy	F1-Score
Stacking Classifier	74.53%	69,29%
Support Vector Machine	69,86%	67,08%
Decision Tree Classifier	64,27%	59.97%
Softmax Regression	57,60%	60,76%
K-Nearest Neighbours	60,40%	51,10%

4 Testing

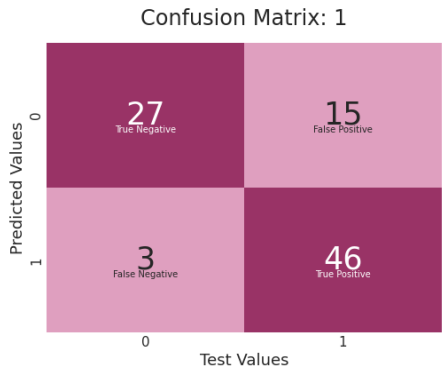
I risultati finali ottenuti sul *testing set* (**91** sample) sono i seguenti:

Classe da predire	Precision	Recall	F1-Score	Contributo
(1) Normal	75%	94%	84%	49
(2) Ischemic changes	55%	67%	60%	9
(3) Old Anterior Myocardial Infraction	60%	100%	75%	3
(4) Old Inferior Myocardial Infraction	67%	67%	67%	3
(5) Sinus tachycardy	50%	33%	40%	3
(6) Sinus bradycardy	75%	60%	67%	5
(7) Ventricular Premature Contraction	0%	0%	0%	1
(9) Left Boundle branch block	100%	100%	100%	2
(10) Right boundle branch block	100%	20%	33%	10
(14) Left Ventricule hypertrophy	0%	0%	0%	1
(15) Atrial Fibrillation or Flutter	0%	0%	0%	1
(16) Other	0%	0%	0%	4
Generale :	69,13%	71,42%	66,48%	91

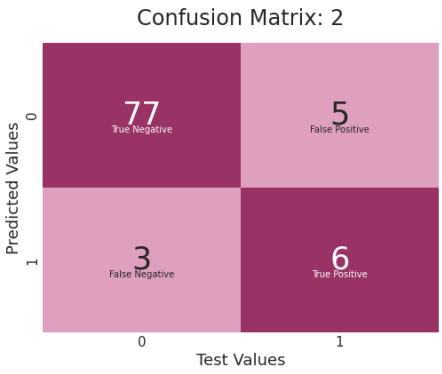
Accuracy : 71,42%

4.1 Confusion matrixs

Le *confusion matrix* sono state realizzate utilizzando la trasformazione **one-vs-rest** tramite la `multilabel_confusion_matrix`, una delle metriche offerta da `sklearn`.
Di seguito sono riportate le *confusion matrix* più rilevanti.

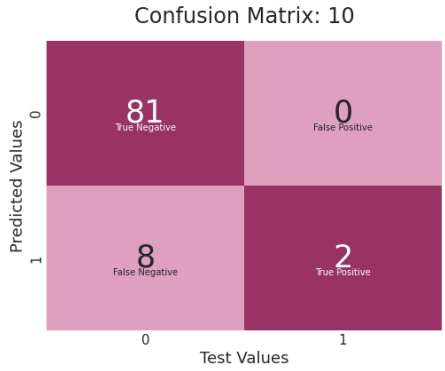


(a) Class: Normal

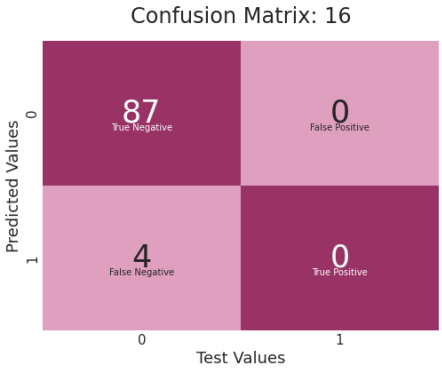


(b) Class: Ischemic changes

-



(c) Class: Right boundle branch block



(d) Class: Other