Università degli studi di Milano-Bicocca

Machine Learning

Progetto Finale

Predicting a Pulsar Star

Autori:
Andrea Corvaglia - 802487 - a.corvaglia@campus.unimib.it
Davide Toniolo - 800458 - d.toniolo2@campus.unimib.it
July 21, 2019

Contents

1 Introduzione **1**
1.1 Stelle di Neutroni e Pulsar 1
1.1.1 Rilevamento 1

2 DataSet **2**
2.1 Statistiche Descrittive 2

3 Modelli **2**

4 Feature selection **3**
4.1 Scelta fra Filtro e Wrapper 3
4.2 Indici di Correlazione 4
4.3 Correlazioni Osservate 4
4.4 Valutazione 4
4.4.1 Cross Classification 5

5 Risultati **6**
5.1 Modelli Completi 6
5.2 Modelli con Features Selection 6

6 Conclusioni **6**

Abstract

Il presente lavoro è volto a risolvere il problema posto nella presentazione del dataset [1], disponibile a questo link.

L'evoluzione tecnologica ha portato ad avere telescopi sempre più potenti ed automatizzati, che generano ad oggi un volume di informazioni che rende necessario l'utilizzo di strumenti per l'analisi automatizzata. In questo lavoro gli autori si prefiggono di realizzare un inducer che sia capace di classificare dei segnali come rumore o segnale, con l'obiettivo di identificare esemplari di pulsar, una categoria particolare di stelle di neutroni. Il problema di classificazione è stato risolto utilizzando degli Artificial Neural Network e una Random Forest, affrontando anche la complicazione delle classi sbilanciate.

1 Introduzione

1.1 Stelle di Neutroni e Pulsar

Le stelle di neutroni sono stelle molto compatte, dense e sede di fenomeni estremi. La nomenclatura non deve trarre in inganno: malgrado vengano chiamate "stelle", sono oggetti molto lontani dalle stelle dell'immaginario comune. Non "bruciano", né sono "infuocate" o molto luminose. Invece sono ciò che rimane del nucleo di una supernova, cioè dell'esplosione di una stella morente.

A differenza delle stelle ordinare, le stelle di neutroni sono formate quasi interamente da neutroni, compressi a una densità 10^{14} volte maggiore della media delle stelle ordinare. Sulla loro superficie ed al loro interno si trovano condizioni estreme, tra cui gravità e campi magnetici estremamente intensi, stati della materia esotici e enormemente densi, etc... . Questa varietà di condizioni particolari ed estreme costituisce un ottimo campo sperimentale per la fisica di base, permettendo potenzialmente di testare la Cromodinamica Quantistica, la Relatività Generale, modelli superfluidi e di avere un unico sistema dove accadono sia fenomeni di Relatività Generale che di Meccanica Quantistica. Il valore sperimentale di queste stelle aumenta di molto se si considera che queste condizioni sono impossibili da ricreare in laboratorio [2].

Purtroppo, le stelle di neutroni sono tanto utili quanto difficili da osservare. A differenza delle stelle ordinarie, esse non "brillano", il che le rende sostanzialmente invisibili se isolate. Nessuno delle stelle che si vedono nel cielo notturno ad occhio nudo è una stella di neutroni, in parte perché emettono poco nel visibile e in parte perché sono troppo deboli.

Uno dei rari casi in cui è possibile rilevare le stelle di neutroni è quando sono molto giovani: dopo l'esplosione di supernova, appena formata la stella di neutroni, essa si trova a ruotare su sé stessa molto velocemente, emettendo, per varie ragioni, due fasci di radiazione elettromagnetica in direzioni antiparallele. L'asse di emissione generalmente non coincide con quello di rotazione, cosicché un osservatore esterno si trova a rilevare (se si trova all'interno del cono di luce) un picco di onde radio, seguito da un ritorno a livelli normali quando la stella continua la sua rotazione. Nel corso del tempo verranno osservati quindi picchi di emissione ad intervalli estremamente regolari, coincidenti con il periodo di rotazione della stella. L'analogia più usata è quella di un faro visto da una nave [3].

Queste giovani stelle di neutroni vengono dette "pulsar", in riferimento alla natura pulsata del segnale rilevato. I modelli teorici stimano che solo una piccola parte di tutte le stelle di neutroni esistenti, circa il 10%, sono pulsar. Il motivo è che virtualmente ogni stella di neutroni è una pulsar all'inizio della sua vita, ma in mancanza di fonti di energia esterne, per via delle forti emissioni, rallenta e smette di emettere dopo circa 1 milione di anni dalla sua nascita, un tempo molto breve su scala astronomica [4]. Dei 100 milioni di stelle di neutroni che si stimano presenti nella Via Lattea, ad oggi solo 2700 sono state scoperte. Per quanto riguarda le altre galassie, ad oggi nessuna stella di neutroni è stata scoperta al di fuori della Via Lattea, in quanto a causa della distanza appaiono troppo deboli.

1.1.1 Rilevamento

Gli astronomi che si occupano di ricercare questi oggetti scandagliano il cielo osservando l'andamento di fondo nella banda dello spettro elettromagnetico prefissata, andando a ricercare gli indicativi picchi di emissione regolari. Le difficoltà presenti in questo processo di ricerca sono numerose, in quanto insieme ai segnali ricercati ne vengono captati anche non di interesse, quali segnali di origine antropica (radio, televisione, cellulari, satelliti) e segnali provenienti da altri oggetti celesti, quali altre stelle, dischi di accrescimento, etc ... . Se le pulsar più vicine a noi e dall'emissione più intensa possono essere rilevate in maniera amatoriale, in quanto

i loro segnali giungono alla Terra sufficientemente intensi da essere distinguibili a occhio nudo nel grafico di un semplice oscilloscopio collegato ad un'antenna, la rilevazione di pulsar più lontane e/o deboli richiede strumenti più sensibili e metodi di analisi più sofisticati.

Vengono utilizzati moderni ed ampi radiotelescopi che raccolgono i segnali provenienti dallo spazio, poi analizzati con una tecnica chiamata analisi spettrale. Le features vengono già estratte dal segnale, secondo criteri definiti in anni di esperienza dagli astronomi specializzati e dagli ingegneri che li affiancano nella progettazione della strumentazione [3], [5].

2 DataSet

Il data set [1] contiene 17,898 osservazioni, di cui 16,259 spurie e 1,639 pulsar. La classificazione è stata fatta manualmente da esperti del settore che forniscono i valori alla variabile target. Le frequenze relative delle due classi sono molto differenti, fattore che è necessario tenere in considerazione nella valutazione delle performance dei modelli. Non sono presenti valori mancanti.

Sono presenti 8 attributi numerici, di cui 2 possono assumere solo valori positivi. Queste variabili fanno riferimento alle proprietà di due distribuzioni di frequenza ottenute dal segnale, dette **integrated profile** e **DM-SNR curve**. Di ciascuna delle due distribuzioni vengono forniti i primi quattro momenti (media, deviazione standard, coefficiente di asimmetria e curtosi). Le variabili nel dataset sono:

1. Mean of the integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.
4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.
9. Class

2.1 Statistiche Descrittive

Creando dei box plot per ogni variabile, si è notato come ciascuna di esse presentasse outlier. Sarà necessario tenerne conto nella fase di training e di test.

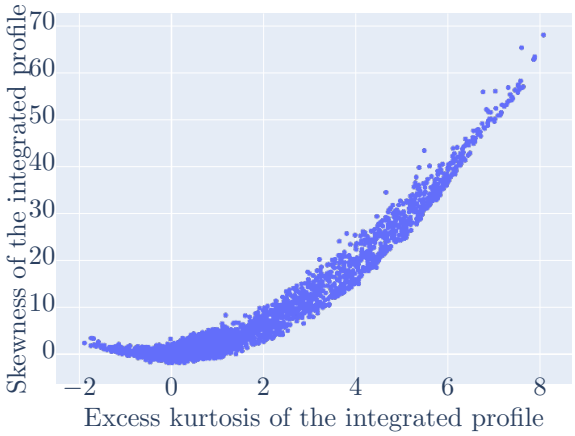

Figure 1: Scatter plot tra l'eccesso di curtosi per il profilo integrato sulle ascisse e la Skewness per la stessa curva sulle ordinate.

Osservando gli scatter plot creati accoppiando le variabili, si nota come molte di esse siano piuttosto legate (es. fig. 1): gli andamenti sono non lineari, ma comunque molto ben definiti. Ne segue che è ragionevole aspettarsi che sia possibile escludere alcune di esse dai modelli ed ottenere comunque una buona performance. Siccome le dipendenze non sono lineari, l'indice di correlazione di Pearson non è necessariamente una buona misura per i legami in questo set di attributi, per cui è necessario usare delle stime differenti che verranno descritte in seguito.

3 Modelli

Il problema è di classificazione binaria. Tra tutti i modelli affrontati nel corso, è possibile escluderne alcuni a priori, facendo considerazioni come: verificare se è adatto alla tipologia di dati ed al loro significato, potere predittivo, costo di allenamento, costo di utilizzo. Vediamoli uno ad uno:

- Logistic Regression: questo modello accetta dati numerici, quindi è applicabile al problema in esame. Tuttavia è molto semplice, quindi è possibile che fatichi a raggiungere alti score nella classificazione.
- Naïve Bayes & Bayesian Network: sono pensati per features discrete e categoriche, mentre questo problema presenta attributi numerici. È possibile discretizzare le variabili ed eseguire comunque l'algoritmo, ma esso comunque non terrebbe conto di proprietà importanti degli attributi quali ordinamento e distanza: potrebbe arrivare ad

impararle, ma parte svantaggiato.

- Random Tree & Random Forest. Essi tengono conto della natura numerica dei dati e possono raggiungere un buon livello di complessità, quindi a priori sono dei buoni candidati. Una Random Forest potrebbe avere più potere predittivo in quanto l'uso di più alberi generalmente riduce la tendenza all'overfitting che si ha con un singolo Random Tree.

- Multilayer Perceptron. Anche questa classe di modelli si applicano a features numeriche, quindi sono utilizzabili nel caso in esame. Possono modellare funzioni molto complesse, quindi hanno abbastanza potere predittivo.

- SVM. Tengono conto della struttura numerica dei dati e generalmente hanno un buon potere predittivo, quindi vale la pena di testarle. Verranno usate con Soft Margins e un Kernel in quanto è improbabile a priori che i dati siano separabili e che non serva un kernel per modellare bene.

In base a queste considerazioni, i modelli Logistic Regression, Naïve Bayes e Bayesian Network sono stati scartati. Si è quindi scelto di testare i rimanenti per capire quale offrisse le performance migliori. Non sono stati implementati modelli di tipo Random Tree, ma solo Random Forest per quanto già discusso sopra. In fase di test, ci si è resi conto che la fase di training di un inducer SVM era estremamente lunga, indipendentemente dal kernel e dagli iperparametri: si è quindi scelto di accantonare le support vector machines in quanto non risultavano apprendibili in pratica. Infine, si è dovuta scegliere l'architettura del Multi-Layer Perceptron, in particolare quanti strati e quanti neuroni per strato. Il primo MLP implementato utilizza i parametri di default, quindi non presenta hidden layer ed a due output neurons, uno per classe. Successivamente, in [3] è stato trovato un riferimento al modello effettivamente utilizzato dagli astronomi per classificare i record: l'architettura è piuttosto semplice, c'è un solo strato nascosto con lo stesso numero di neuroni dello strato di input. Si è scelto allora di implementare anche questo modello, senza sostituire l'ANN precedente, per andare a comparare il potere predittivo delle due architetture.

Che differenze di performance ci si può aspettare, a priori, tra Random Forest e ANN? Negli ultimi anni, i NN hanno avuti molti successi nel campo del machine learning. Tuttavia, essi sono stati ottenuti soprattutto in problemi non strutturati, come quelli di computer vision e di riconoscimento vocale. Infatti, uno dei grandi vantaggi degli ANN rispetto agli altri metodi di classificazione è che essi hanno la possibilità di scegliere, estrarre e trasformare le features più rilevanti dai dati in modo automatico, mentre per gli altri algoritmi è spesso richiesta una fase di feature selection, extraction and transformation manuale precedente all'apprendimento del modello [6].

Tuttavia nel problema in esame non viene fornito il segnarle originario, ma è già stato effettuato un grosso lavoro di feature extraction ed, in parte, selection. Quindi, pur non essendo possibile determinare a priori quale sarà l'algoritmo più performante, è possibile prevedere che gli ANN non avranno quel ruolo di tecnologia abilitante che hanno avuto nel campo della percezione sensoriale delle macchine. In più, il loro eventuale vantaggio prestazionale si riduce se rapportato al maggiore costo computazionale del loro apprendimento.

4 Feature selection

Spesso in una classificazione non tutte le variabili sono utili o necessarie per ottenere un modello con buone capacità predittive. A parità di risultati, utilizzare meno attributi, fornisce un vantaggio in termini di costo computazionale sia in fase di apprendimento e sia in fase di predizione. Inoltre aumenta la comprensibilità del modello, rendendo possibile l'estrazione di nuova conoscenza dal modello da parte degli esperti, invece di utilizzarlo solamente come una "scatola nera" per le previsioni.

4.1 Scelta fra Filtro e Wrapper

Per scegliere quali features eliminare, si è scelto di utilizzare un filtro per tre ragioni. La prima è che già in fase di analisi esplorativa ci si è resi conto che molte coppie di variabili sono fortemente legate fra loro. Se, per esempio, la "Skewedness of the integrated profile" è una funzione polinomiale della "Excess kurtosis of the integrated profile" (fig. 1), è ragionevole aspettarsi che un buon modello riesca a predire efficacemente le classi anche omettendo una delle due. È quindi possibile rimuovere alcune variabili a priori, senza valutare ed apprendere un'istanza di un modello, semplicemente osservando le correlazioni tra le variabili. Utilizzando quindi opportune misure di interdipendenza fra gli attributi si può implementare un filtro che seleziona quali features utilizzare.

La seconda ragione è puramente pratica: utilizzare un wrapper invece che un filtro ha un

costo computazionale più alto, dovuto al fatto che il modello va appreso ad ogni iterazione dell'algoritmo che implementa il wrapper. I filtri, invece, non tengono conto del modello, ma utilizzano criteri a priori per la selezione delle variabili, risultando in algoritmi più snelli.

Infine, il terzo motivo riguarda la maggiore comprensibilità dei risultati di un algoritmo di filtro rispetto a quelli di uno di wrapper. Infatti se quest'ultimo esclude una variabile, può essere difficile comprendere se essa è stata rimossa perché non è abbastanza influente sull'attributo di classe o perché è molto correlata ad un'altra feature già inclusa. Un filtro invece tiene conto solo dei legami tra gli attributi.

4.2 Indici di Correlazione

Come già accennato in precedenza, l'indice di correlazione di Pearson non è adatta a stimare l'interdipendenza tra questi attributi, in quanto le dipendenze funzionali tra le variabili non sono lineari. Tuttavia, osservando gli scatter plot, si nota che tutte le relazioni sono (quasi) monotone: è quindi possibile utilizzare con efficacia due misure:

1. Spearman's ρ. Valuta quanto fortemente ci sia una dipendenza monotona tra due variabili. In mancanza di valori ripetuti, un valore di $+1$ indica una perfetta monotonia crescente, mentre un valore di -1 una decrescente.

2. Kendall's τ. Valuta quanto le coppie di valori delle due variabili rispettino l'ordinamento dei quantili complessivi delle due variabili. Ha valori in $[-1; 1]$ ed ha un'interpretazione simile alla Spearman's ρ.

Questi due indici sono entrambi non parametrici e fanno parte dei cosiddetti *coefficienti di correlazione per ranghi* [7].

4.3 Correlazioni Osservate

Costruendo le matrici contenenti i due indici scelti (tabelle 3 e 4), si può apprezzare come i valori lì riportati siano in accordo con quanto suggerito dagli scatter plot: i momenti di ogni curva sono collegati tra loro, ad eccezione delle σ. Le variabili più legate al target sono, in ordine: eccesso di curtosi del profilo integrato, media e skewness dello stesso, media e varianza DM-SNR e così via.

Rimane quindi da capire quali variabili siano rilevanti per la classificazione e quali non lo siano. Verranno quindi provati due approcci. Il primo è quello di fornire una delle due misure di correlazione tramite il nodo **Rank Correlation** al nodo **Correlation Filter** ed analizzare quante e quali variabili vengano incluse a variare della soglia, per poi andare ad apprendere un'istanza di un modello Random Forest ed osservare quanto bene riesca a prevedere la classe.

Il secondo è quello di scegliere manualmente le variabili sulla base della matrice di correlazione (con la variabile target inclusa) calcolata secondo uno dei due indici, secondo uno pseudoalgoritmo di forward inclusion:

1. Si decide quante variabili si desidera utilizzare nel modello.

2. La variabile più correlata con target viene inclusa.

3. Si prende la seconda variabile più correlata con il target e se è debolmente legata a quelle già presenti si include. Altrimenti, la si esclude e si passa alla terza, e così via fino a raggiungere il numero di variabili desiderato o fino a che si scartano tutte quelle rimanenti.

Anche in questo secondo caso, una volta selezionate le features si passa alla fase di apprendimento di un'istanza di un modello di Random Forest e alla sua valutazione. I risultati saranno discussi più in basso, nel opportuna sezione.

4.4 Valutazione

Un'analisi della classificazione consiste nello sviluppo di modelli di classificazione di diverso tipo o con diversi parametri, con lo scopo di selezionare una specifica formulazione del modello di classificazione in grado di assicurare la miglior accuratezza predittiva possibile. Un modello è giudicato migliore sulla base di diverse caratteristiche, la principale delle quali è l'efficacia predittiva, che normalmente viene valutata tramite l'accuratezza (classificazioni corrette su numero totale di classificazioni). Dovendo fare una classificazione su un attributo il cui esito positivo è raro, l'accuratezza diventa un approccio fuorviante, perché tiene conto nello stesso modo delle due classi. Un esempio eclatante è l'approccio *0 Rule*, che consiste nell'assegnare a tutti i record il valore della classe dominante: con il dataset in esame otterrebbe un'accuratezza di circa il 90 %, pur essendo completamente inutile per uno scopo predittivo. Diventa quindi necessario utilizzare delle metriche diverse per la valutazione dei risultati, che prediligano la performance sulla

classe rara. Le due metriche naturali per questo proble ma sono dunque:

$$Recall = \frac{VeriPositivi}{VeriPositivi + FalsiNegativi}$$

$$Precisione = \frac{VeriPositivi}{VeriPositivi + FalsiPositivi}$$

Dove i veri positivi sono i record classificati correttamente come pulsar, mentre i falsi negativi sono delle pulsar classificate erroneamente come spurie. Il Recall è dunque la frazione di pulsar effettivamente classificate come tali, mentre la precisione è la frazione di pulsar reali tra i record classificati positivamente. La valutazione dei classificatori nei termini delle metriche appena descritte è stata ottenuta tramite l'utilizzo del nodo **Score** applicato ai risultati della classificazione.

4.4.1 Cross Classification

Una volta che un'istanza del modello è stata appresa, è buona norma valutarne le capacità predittive su dei dati non presenti nel training set per ridurre il rischio di overfitting. Avendo a disposizione una quantità limitata di dati, si preferisce non utilizzarli tutti per apprendere l'istanza del modello, ma si partiziona il dataset in un train set e un test set (Hold Out).

Siccome nel caso in questione il dataset presenta molti outlier, la valutazione su un singolo dataset potrebbe essere fuorviata da queste osservazioni insolite: per avere una stima più robusta delle capacità predittive dell'istanza, ed indipendente dal particolare test set, sono necessarie tecniche di valutazione più robuste, come la cross validation.

Essa consiste nel creare una partizione dell'intero dataset in K sotto dataset con approssimativamente la stessa numerosità. Ciascuno di questi sottoinsiemi a turno è tenuto come test set, mentre i restanti nove compongono il training set. Il modello viene appreso K volte sui K training set e valutato sul training set corrispondente. Nel caso in esame, come detto nella descrizione del dataset, sono presenti diversi outlier. Questo rende rischioso avere test set con una distribuzione arbitraria dei record, perché potrebbero non essere rappresentativi del dataset, ma essere formati da una percentuale più elevata di outlier e quindi portare ad una valutazione fuorviante. Siccome, però, nella cross validation ogni osservazione compare una ed una sola volta nel test set, questo assicura che osservazioni particolarmenti influenti sullo score finale non possano ingannare la valutazione.

Inoltre è importante notare che la partizione è stata creata con un sampling stratificato rispetto alla variabile di classe, in modo da mantenere nei train sets una frequenza relativa simile all'originale.

Il metodo è stato implementato in KNIME utilizzando i nodi **X-partitioner** e **X-aggregator** che generano un ciclo che esegue le fasi della Cross Validation, restituendo la matrice originaria con aggiunta una colonna con la classe predetta ed una con l'id del fold in cui quella osservazione ha fatto parte del test set.

A questo punto si vogliono confrontare le performance dei diversi modelli, per capire se le differenze negli score sono significative. Viene quindi effettuata una One-Way Balanced ANOVA, per la quale tuttavia è necessario sapere recall e precisione per ogni fold di ciascun modello, cosa che il nodo **Scorer** non permette. È quindi necessario trasformare manualmente la tabella restituita da **X aggregator** per ottenerne una che contenga una riga per fold con i valori di recall e precisione *specifici per quel fold*. Infine, una volta creata questa tabella per ogni modello, vanno aggregate in un'unica matrice.

Il procedimento dettagliato è il seguente:

1. viene aggiunta una colonna contenente una tra le seguenti stringhe di testo: "TP", "FP", "TN" o "FN", scelta opportunamente.

2. questa nuova colonna viene binarizzata, cioè trasformata in quattro nuove colonne, una per ogni valore possibile. Esse contengono dei valori binari, per indicare a quale delle quattro categorie corrisponde ogni osservazione.

3. vengono aggregate le righe per ogni fold, in modo da avere il numero di veri positivi, ec... totali per ogni fold.

4. vengono calcolati recall e precisione per ogni fold.

Inoltre, si è deciso di sfruttare il sample di valutazioni ottenuto dalla validazione per ricavare una stima del valore vero del recall e della precisione utilizzando la distribuzione t di Student con un valore di confidenza del 95%, rappresentando poi gli intervalli tramite boxplot. Partendo dai sample delle valutazioni delle metriche per ogni classificatore utilizzato, si è proseguito con un test di ipotesi statistico in particolare un ANOVA ad un via, la cui ipotesi nulla consiste nell'uguaglianza tra le prestazioni dei classificatori.

5 Risultati

5.1 Modelli Completi

L'ANOVA effettuata sui modelli completi (Tabelle 1, 2) non evidenzia differenze significative né in recall, né in precisione. I tre modelli analizzati si possono considerare dunque equivalenti dal punto di vista del potere predittivo. La rete neurale più complessa non fornisce a questo livello risultati migliori di quella meno profonda. Inoltre la Random Forest dà risultati equivalenti in score ai due NN. Siccome essa è più rapida sia in fase di apprendimento sia durante le previsioni, dopo le analisi fatte risulta preferibile alle reti neurali artificiali. I modelli appresi successivamente, cioè quelli con le features filtrate sono stati solo Random Forest.

5.2 Modelli con Features Selection

Inizialmente è stato valutato il potere predittivo dei modelli con le features selezionate dal nodo **Correlation Filter**, al variare del parametro soglia.

Successivamente sono state selezionate le variabili sulla base della Spearman's ρ e tramite lo pseudoalgoritmo descritto precedentemente. Va evidenziato che le variabili selezionate coi due metodi non coincidono, in quanto il filtro non tiene conto dell'importanza delle features, ma solo della correlazione tra queste. Di conseguenza a parità di numero di attributi inclusi, si è arrivati ad avere valori di recall e precision migliori di quelli ottenuti dalle istanze dei modelli filtrati, capendo quali fossero le variabili fondamentali per la previsione e quali solo accessorie.

Excess curtosis of integrated profile risulta di gran lunga la più importante: un modello che includa solamente questa variabile giunge comunque ad ottenere un valore di recall di 0.79. La ragione è facile da comprendere andando a vedere le regole scelte dai random trees: per ECOIP c'è una separazione tra i record negativi e quelli positivi. I primi hanno quasi tutti valori al di sotto di $1.0 - 1.1$, mentre i secondi quasi tutti al di sopra. Nessun'altra variabile presenta una distinzione così netta.

Seconde in importanza sono le due skewness, che se incluse portano ad una recall di 0.84. Aggiungendo ulteriori attributi, si apportano solo piccole correzioni agli score, fenomeno ben spiegato dal fatto che le 3 variabili già incluse siano molto correlate con tutte le altre, quindi se si utilizzassero anche le rimanenti 5 si avrebbero poche informazioni aggiuntive.

È necessario sottolineare come i modelli semplificati non siano necessariamente preferibili come inducer, in quanto nella pratica gli astronomi preferiscono avere valori di recall più alti possibile [5].

6 Conclusioni

Dai risultati del test ANOVA risulta che nessuno dei modelli è significativamente migliore rispetto agli altri. Si preferisce quindi utilizzare le Random Forest per questioni di costo computazionale.

L'analisi della correlazione evidenzia una netta gerarchia di importanza tra le variabili, nonché una palese dipendenza tra alcune di esse. Queste informazioni hanno permesso di ridurre gli attributi utilizzati a tre, pur mantenendo quasi inalterata la capacità di classificazione, ottenendo quindi un modello molto più semplice ed interpretabile. In ogni caso le performance ottenute sono buone, ma non bastano per delle serie applicazioni in astronomia, dove sono richiesti valori di recall molto più vicini ad 1 [5].

Sulla base di quanto detto i possibili miglioramenti sono:

- Utilizzare un dataset con più informazioni, per esempio con un maggior numero di variabili tra cui scegliere.
- Indirizzare il lavoro degli algoritmi verso una maggiore attenzione per la classe rara tramite l'aggiunta di una matrice di score.
- Implementare una iperarametrizzazione automatica per i modelli.

References

[1] "Kaggle predicting pulsar star," https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star.

[2] D. Maoz, Astrophysics in a Nutshell. Princeton university press, 2016, vol. 16.

[3] D. Thornton, "The high time resolution radio sky," Ph.D. dissertation, The University of Manchester (United Kingdom), 2013.

[4] M. Camenzind, Compact objects in astrophysics. Springer, 2007.

[5] R. J. Lyon, B. Stappers, S. Cooper, J. Brooke, and J. Knowles, "Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach," Monthly Notices of the

Model	TP	FP	TN	FN	Recall	Precisione	F-measure	Accuracy
RF	1376	103	16156	263	0.84	0.93	0.84	0.88
Simple MLP	1397	117	16142	242	0.85	0.92	0.89	0.98
Complex MLP	1407	128	16131	232	0.86	0.92	0.89	0.98

Table 1: Punteggi dei modelli completi dopo la cross validation.

Score	Fonte	SS	df	SS Media (10^{-4})	F	p
Recall	Between Groups	0.001869	2	9.35	2.31	0.11
Recall	Within Groups	0.0108	27	3.98	-	-
Recall	Total	0.0126	29	-	-	-
Precision	Between Groups	0.0009	2	4.59	0.89	0.42
Precision	Within Groups	0.0138	27	5.13	-	-
Precision	Total	0.0148	29	-	-	-

Table 2: Risultati dell'ANOVA

Kendall's tau rank correlation coefficient									
	μ IP	σ IP	κ IP	γ IP	μ DM	σ DM	κ DM	γ DM	**Target**
μ IP	1.000	0.350	-0.706	-0.466	-0.050	-0.059	0.054	0.056	-0.368
σ IP	0.350	1.000	-0.350	-0.713	0.006	-0.002	0.001	0.003	-0.255
κ IP	-0.706	-0.350	1.000	0.486	0.056	0.066	-0.058	-0.061	0.385
γ IP	-0.466	-0.713	0.486	1.000	0.041	0.050	-0.046	-0.048	0.360
μ DM	-0.050	0.006	0.056	0.041	1.000	0.818	-0.925	-0.902	0.320
σ DM	-0.059	-0.002	0.066	0.050	0.818	1.000	-0.809	-0.864	0.322
κ DM	0.054	0.001	-0.058	-0.046	-0.925	-0.809	1.000	0.939	-0.314
γ DM	0.056	0.003	-0.061	-0.048	-0.902	-0.864	0.939	1.000	-0.316
Target	-0.368	-0.255	0.385	0.360	0.320	0.322	-0.314	-0.316	1.000

Table 3: Matrice di correlazione tra le variabili, con coefficiente di correlazione Kendall's Tau. IP indica gli attributi relativi all'integrated profile, mentre DM sta per la curva DM-SNR. Gli attributi sono indicati con μ per la media, σ per la deviazione standard, κ per la curtosi e γ per la skewness.

Spearman's rank correlation coefficient									
	μ IP	σ IP	κ IP	γ IP	μ DM	σ DM	κ DM	γ DM	**Target**
μ IP	1.000	0.499	-0.880	-0.635	-0.077	-0.090	0.081	0.085 -	0.451
σ IP	0.499	1.000	-0.496	-0.876	0.007	-0.005	0.003	0.005	-0.313
κ IP	-0.880	-0.496	1.000	0.657	0.087	0.100	-0.089	-0.093	0.472
γ IP	-0.635	-0.876	0.657	1.000	0.065	0.078	-0.072	-0.075	0.441
μ DM	-0.077	0.007	0.087	0.065	1.000	0.950	-0.991	-0.986	0.393
σ DM	-0.090	-0.005	0.100	0.078	0.950	1.000	-0.946	-0.972	0.394
κ DM	0.081	0.003	-0.089	-0.072	-0.991	-0.946	1.000	0.994	-0.384
γ DM	0.085	0.005	-0.093	-0.075	-0.986	-0.972	0.994	1.000	-0.386
Target	-0.451	-0.313	0.472	0.441	0.393	0.394	-0.384	-0.386	1.000

Table 4: Matrice di correlazione tra le variabili, con coefficiente di correlazionedi Spearsman. IP indica gli attributi relativi all'integrated profile, mentre DM sta per la curva DM-SNR. Gli attributi sono indicati con μ per la media, σ per la deviazione standard, κ per la curtosi e γ per la skewness.

Royal Astronomical Society, vol. 459, no. 1, pp. 1104–1123, 2016.

[6] C. Francois, "Deep learning with python," 2017.

[7] "Rank correlation," https://en.wikipedia.org/wiki/Rank_correlation#General_correlation_coefficient.