PROGETTO DATAMINING:

ANALISI SULLE PULSAR

Agosto 2019

Riccardo Soro

INDICE:

1 Analisi del Dominio

In questa sezione verranno effettuate delle analisi sui dati del Dataset prima di apportarne delle modifiche, al fine di capirne la composizione e di poter comprendere gli attributi dai quali sono composti i record all'interno del Dataset.

1.1 Descrizione del Dominio

I dati all'interno del Dataset analizzato appartengono a delle stelle e l'obiettivo è quello di capire se la stella analizzata è una Pulsar oppure no.

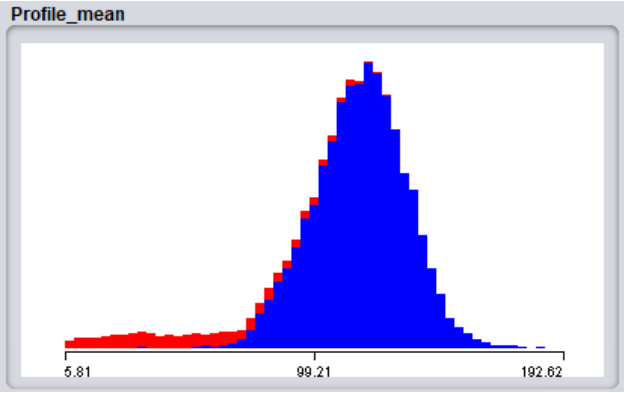
Una pulsar, nome che stava originariamente per sorgente radio pulsante, è una stella di neutroni. Nelle prime fasi della sua formazione, in cui ruota molto velocemente, la sua radiazione elettromagnetica in coni ristretti è osservata come impulsi emessi ad intervalli estremamente regolari.

Esse sono rare da trovare e sono molto utili poiché vengo utilizzate per calcolare la posizione di altri corpi celesti in movimento. Ogni Pulsar ha delle proprie caratteristiche, le quali messe insieme fungono da impronta digitale per essa; in questo modo è possibile distinguerle in maniera univoca. Queste caratteristiche sono il segnale elettromagnetico emesso, il suo periodo di rotazione e la quantità di elettroni liberi tra la stella e l'osservatore.

1.2 Descrizione del Dataset

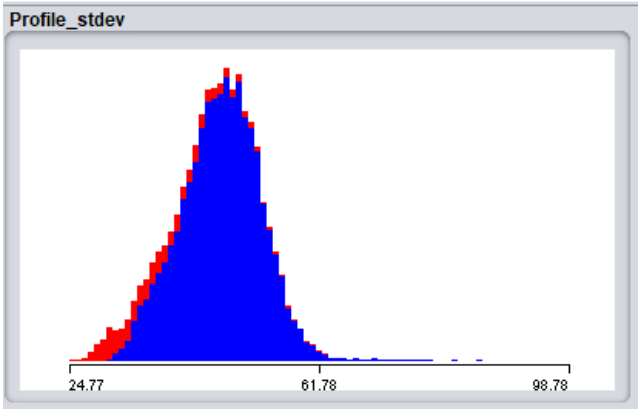
Il Dataset è disponibile al seguente link:
https://archive.ics.uci.edu/ml/datasets/HTRU2
Nel Dataset sono presenti 17898 elementi composti da 9 attributi ciascuno:

1) Profile_mean: Rappresenta la media della firma della pulsar.



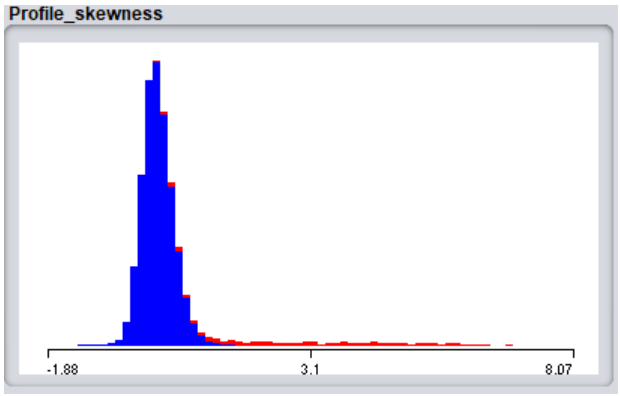
Media: 100,08 Deviazione Standard: 25,65

2) Profile_stdev: Rappresenta la deviazione standard della firma della pulsar.



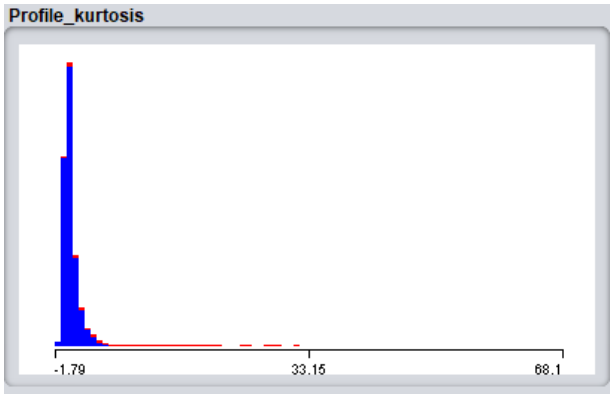
Media: 46,55 Deviazione Standard: 6,84

3) Profile_skewness: Rappresenta l'indice di asimmetria della firma della pulsar.



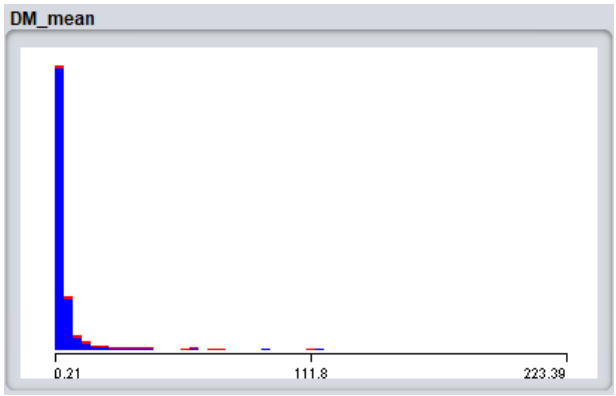
Media: 0,48 Deviazione Standard: 1,06

4) Profile_kurtois: Rappresenta l'indice di Curtosi della firma della pulsar.



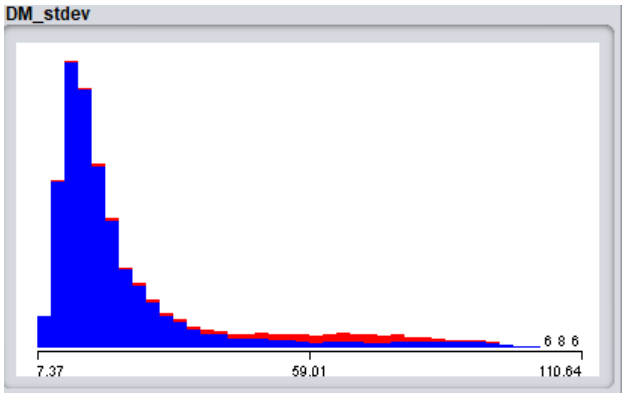
Media: 1,77 Deviazione Standard: 6,17

5) DM_mean: Rappresenta la media degli elettroni liberi.



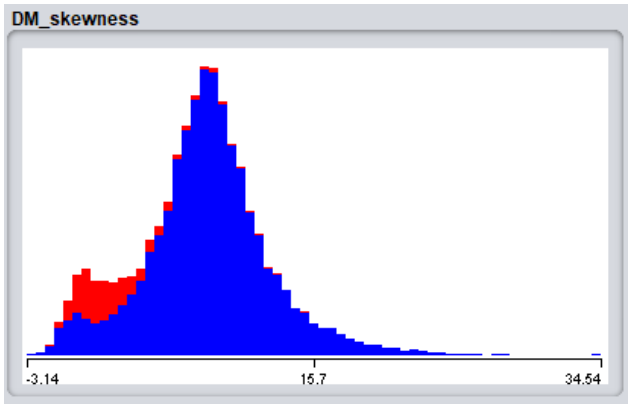
Media: 12,61 Deviazione Standard: 29,47

6) DM_stdev: Rappresenta la deviazione standard degli elettroni liberi.



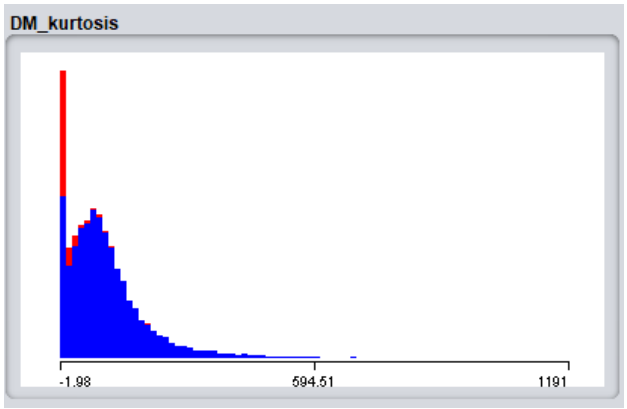
Media: 26,33 Deviazione Standard: 19,47

7) DM_skewness: Rappresenta l'indice di asimmetria degli elettroni liberi.



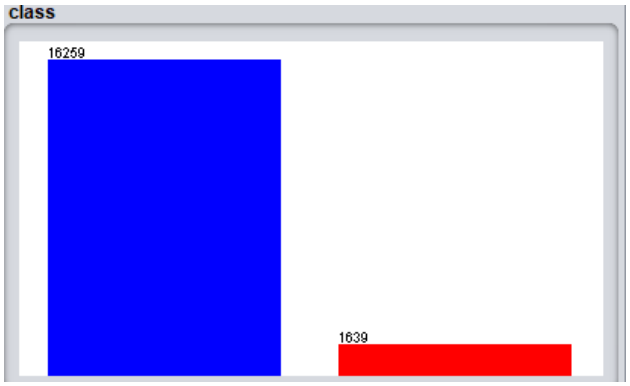
Media: 8,30 Deviazione Standard: 4,50

8) DM_kurtosis: Rappresenta l'indice di Curtosi degli elettoni liberi.



Media: 104,86 Deviazione Standard: 106,52

9) Class: Rappresenta la classe di appartenenza della stella.



Tutti gli attributi sono continui ed il dataset risulta sbilanciato con una media centrata sulla classe non-pulsar.

2 Preprocessing dei Dati

In questa sezione vengono effettuate delle modifiche ai dati al fine di aumentare la qualità dell'output dei vari algoritmi e di garantire delle performance migliori.
I valori degli attributi sono numerici continui e con i range molto differenti tra loro, sarà perciò utile applicare tecniche di trasformazione e discretizzazione.
Non si è ritenuto necessario aggregare o eliminare attributi in quanto il numero di questi è molto ridotto e non sono presenti attributi linearmente dipendenti tra di loro.

2.1 Pulizia

In questa fase si è controllato il Dataset al fine di rimuovere eventuali dati duplicati e gestire i dati con attributi vuoti.
Come scritto nella descrizione del Dataset sul link sopra riportato i dati risultano controllati manualmente da degli esperti, quindi è stata esclusa la presenza di errori e di eccessivo rumore.

È stata comunque svolta una fase di controllo, ma non è stato individuato alcun record duplicato né la presenza di valori mancanti.

2.2 Normalizzazione

Come precedentemente sottolineato, i range degli attributi sono estremamente diversi, è stato quindi deciso di applicare il filtro non supervisionato *Standardize* con i parametri di default.
Questo filtro permette di manipolare i dati del dataset al fine di ottenere per ogni attributo la media dei suoi valori di 0 e la deviazione standard di 1.
Questo filtro risolve il problema del range elevato, il quale disturba e rende inefficaci alcuni algoritmi che si basano sulla distanza (KNN).

2.3 Campionamento

Essendo il Dataset fortemente sbilanciato, è stata presa la decisione di campionarlo per ottenere un dataset più omogeneo al fine di aumentare la qualità dei risultati di quegli algoritmi specializzati nel lavorare su dataset omogenei.
Di conseguenza è stato generato un dataset di questo tipo per testare i risultati degli algoritmi che più si prestano a dataset di questo tipo.

2.4 Discretizzazione

Alcuni algoritmi, come i Bayesiani, ottengono risultati migliori discretizzando il dataset in quanto si basano sulle sulla distribuzione delle probabilità.
E' stato perciò deciso di creare un dataset discretizzato sul quale testare questi algoritmi.

3 Classificatori

3.1 KNN

Per questo algoritmo è stato utilizzato il dataset standardizzato, è stato eseguito anche sul database iniziale ma, come previsto, i risultati sono stati peggiori.

```
Correctly Classified Instances       17517               97.8713 %
Incorrectly Classified Instances       381                2.1287 %
Kappa statistic                          0.8652
Mean absolute error                      0.0317
Root mean squared error                  0.1364
Relative absolute error                 19.0376 %
Root relative squared error             47.2886 %
Total Number of Instances            17898

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,994    0,173    0,983      0,994    0,988      0,867    0,963     0,994     0
                 0,827    0,006    0,933      0,827    0,877      0,867    0,963     0,906     1
Weighted Avg.    0,979    0,157    0,978      0,979    0,978      0,867    0,963     0,986

=== Confusion Matrix ===

     a     b   <-- classified as
 16161    98 |     a = 0
   283  1356 |     b = 1
```

Dopo vari tentativi il risultato migliore è stato raggiunto con K=6 e il peso della distanza settato a 1/distanza eseguito sul database standardizzato.

3.2 Decision Tree

Per questo algoritmo è stato utilizzato il dataset standardizzato e dopo vari tentativi il parametro migliore per il confidenceFactor è stato stimato essere 0.2

```
Correctly Classified Instances       17516               97.8657 %
Incorrectly Classified Instances       382                2.1343 %
Kappa statistic                          0.8663
Mean absolute error                      0.034
Root mean squared error                  0.1364
Relative absolute error                 20.4113 %
Root relative squared error             47.3021 %
Total Number of Instances            17898

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,993    0,162    0,984      0,993    0,988      0,867    0,955     0,990     0
                 0,838    0,007    0,922      0,838    0,878      0,867    0,955     0,871     1
Weighted Avg.    0,979    0,148    0,978      0,979    0,978      0,867    0,955     0,980

=== Confusion Matrix ===

     a     b   <-- classified as
 16142   117 |     a = 0
   265  1374 |     b = 1
```

La precisione è molto alta, ma l'algoritmo ha un livello di recall della classe pulsar un po' basso, di conseguenza si è testato l'algoritmo anche sul dataset reso bilanciato con diversi risultati.

```
Correctly Classified Instances        3050               93.5583 %
Incorrectly Classified Instances       210                6.4417 %
Kappa statistic                          0.8712
Mean absolute error                      0.0988
Root mean squared error                  0.2339
Relative absolute error                 19.7569 %
Root relative squared error             46.7896 %
Total Number of Instances             3260

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,967    0,096    0,910      0,967    0,938      0,873    0,960     0,937     0
                 0,904    0,033    0,965      0,904    0,934      0,873    0,960     0,956     1
Weighted Avg.    0,936    0,064    0,937      0,936    0,936      0,873    0,960     0,946

=== Confusion Matrix ===

    a    b   <-- classified as
 1576   54 |    a = 0
  156 1474 |    b = 1
```

Durante questa prova il confidenceFactor è stato settato a 0.3, sono stati testati anche altri valori ma il risultato migliore è stato raggiunto con questo.

Col dataset bilanciato la recall che prima era scarsa ora risulta accettabile.

3.3 Naive Bayes

Per questo algoritmo è stato utilizzato il dataset iniziale e impostando useKernelEstimator a True. Sono state fatte prove sia sul dataset bilanciato che sbilanciato e la precisione maggiore si raggiunge con quello sbilanciato.

```
Correctly Classified Instances       17365               97.022  %
Incorrectly Classified Instances       533                2.978  %
Kappa statistic                          0.824
Mean absolute error                      0.0322
Root mean squared error                  0.1621
Relative absolute error                 19.3654 %
Root relative squared error             56.2072 %
Total Number of Instances            17898

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,982    0,144    0,985      0,982    0,984      0,824    0,966     0,995     0
                 0,856    0,018    0,825      0,856    0,840      0,824    0,966     0,899     1
Weighted Avg.    0,970    0,132    0,971      0,970    0,970      0,824    0,966     0,986

=== Confusion Matrix ===

     a     b   <-- classified as
 15962   297 |     a = 0
   236  1403 |     b = 1
```

3.4 Rules (JRIP)

Anche in questo caso usando il dataset bilanciato otteniamo una precisione minore ma una recal per le pulsar nettamente superiore.

Sono stati fatti molti tentativi per trovare la combinazione di parametri che permettessero la massima qualità di risultato.

```
Correctly Classified Instances       17507               97.8154 %
Incorrectly Classified Instances       391                2.1846 %
Kappa statistic                          0.8653
Mean absolute error                      0.0366
Root mean squared error                  0.1406
Relative absolute error                 21.9715 %
Root relative squared error             48.7488 %
Total Number of Instances            17898

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
                 0,991    0,148    0,985      0,991    0,988      0,866    0,926
                 0,852    0,009    0,904      0,852    0,877      0,866    0,926
Weighted Avg.    0,978    0,135    0,978      0,978    0,978      0,866    0,926

=== Confusion Matrix ===

     a     b   <-- classified as
 16110   149 |     a = 0
   242  1397 |     b = 1
```

In questa figura sono presenti i dati proveniente dall'esecuzione dell'algoritmo sul dataset sbilanciato e con i seguenti parametri: -F 3 -N 3.0 -O 5 -S 1

```
Correctly Classified Instances        3064               93.9877 %
Incorrectly Classified Instances       196                6.0123 %
Kappa statistic                          0.8798
Mean absolute error                      0.1015
Root mean squared error                  0.2324
Relative absolute error                 20.2968 %
Root relative squared error             46.4828 %
Total Number of Instances             3260

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
                 0,972    0,093    0,913      0,972    0,942      0,882    0,957
                 0,907    0,028    0,970      0,907    0,938      0,882    0,957
Weighted Avg.    0,940    0,060    0,942      0,940    0,940      0,882    0,957

=== Confusion Matrix ===

    a    b   <-- classified as
 1585   45 |    a = 0
  151 1479 |    b = 1
```

In questa immagine invece si possono osservare i risultati dell'algoritmo applicato al dataset bilanciato e con i seguenti parametri: -F 3 -N 3.0 -O 2 -S 1

Possiamo quindi notare come col dataset sbilanciato la precisione sia ottima a discapito però della recal per la classe pulsar, che invece risulta essere accettabile nella seconda immagine a discapito della precisione.

4 Conclusione

Il classificatore che ha portato ai migliori risultati è JRIP, il quale massimizza la recal per la classe pulsar mantenendo comunque una buona precisione.
In conclusione, posso dire che il dataset ha dimostrato una complessità non prevista in quanto l'argomento trattato è complicato ed si è resa necessaria una fase di studio del dominio che ha impiegato abbastanza tempo.
La fase di analisi dei dati è stata molto semplice poiché tutti i dati erano della stessa tipologia e attraverso il grafico a dispersione generato automaticamente da Weka ho potuto avere una visione dei dati molto intuitiva, inoltre il dataset risultava pulito e completo e ciò ha facilitato notevolmente il lavoro.
Ho trovato molto utile questo progetto in quanto mi ha permesso di mettere in pratica concetti teorici studiati per il corso e mi ha costretto a fare uno studio su un dominio che mi ha sempre affascinato.