Analisi del dataset KDD99

versione: kdd_10_percent_red_CFS_BFSsmall10

Analisi degli attributi

- **Protocol**: indica il tipo di protocollo utilizzato nella connessione
- **Service:** indica il tipo di servizio destinatario
- **Flag:** eventuale valori dei campi flag del pacchetto, possono dare informazioni sullo stato della connessione (*e quindi anche se potenzialmente normale o no*)
- **src_byte:** la quantità di byte trasferita dal mittente al destinatario in una singola connessione
- **dst_byte:** la quantità di byte trasferita dal destinatario al mittente in una singola connessione
- **Land binarized:** se l' IP e i numeri di porta (*del pacchetto*) del mittente e del destinatario sono uguali, allora vale 1, altrimenti 0
- **Wrong fragment:** numero totale di frammenti sbagliati
- **Urgent:** numero di pacchetti con il bit "*Urgent*" settato in questa connessione
- **Logged_in_binarized:** status del login, 1 se è loggato con successo, 0 altrimenti
- **count:** numero di connessioni allo stesso host negli ultimi due secondi
- **srv_count:** numero di connessioni alla stessa porta (stesso servizio) negli ultimi due secondi
- **same_srv_rate:** la percentuale delle connessioni rivolte allo *stesso* servizio, nelle connessioni studiate dall' attributo **count**
- **diff_srv_rate:** la percentuale delle connessioni rivolte a *diversi* servizi, nelle connessioni studiate dall' attributo **count**
- **dst_host_same_srv_rate:** la percentuale delle connessioni rivolte a diversi servizi, fra le connessioni studiate dall' attributo **dst_host_count** (ma questo attributo non è presente fra quelli selezionati dall' algoritmo)
- **dst_host_same_src_port_rate:** la percentuale delle connessioni rivolte allo stesso servizio (*stessa porta*), fra quelle studiate dall' attributo **dst_host_srv_count** (ma questo attributo non è presente fra quelli selezionati dall' algoritmo)
- **Class:** possibili classi utilizzati dal classificatore

Domanda: non sarebbe opportuno aggiungere anche i due attributi mancanti? Oppure no?

Studio dell' accuratezza di base (baseline accuracy)

Lo studio è stato svolto utilizzando l'algoritmo **ZeroR** e come opzione per il test il *training set* stesso.

```
=== Summary ===

Correctly Classified Instances       28079               56.847  %
Incorrectly Classified Instances     21315               43.153  %
Kappa statistic                          0
Mean absolute error                      0.0695
Root mean squared error                  0.1864
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances            49394
```

Accuratezza di base: 56.85%

Studio del dataset con J48 (con impostazioni di deafult)

Per questa prova, come per le altre, è stato scelto come metodo di Test la Cross-validation con valore 10.

(il risultato del test è possibile visionarlo sul file `J48_defaultConfig`*)*

Accuratezza

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       49323               99.8563 %
Incorrectly Classified Instances        71                0.1437 %
Kappa statistic                          0.9976
Mean absolute error                      0.0003
Root mean squared error                  0.0127
Relative absolute error                  0.3741 %
Root relative squared error              6.8353 %
Total Number of Instances            49394

=== Detailed Accuracy By Class ===
```

L' accuratezza data risulta essere estremamente ottima, circa il 99.86%, molto più alta rispetto a quella di base.
La precisione e l'affidabilità del classificatore vengono confermati anche dagli altri valori, descritti successivamente alla percentuale di accuratezza.

Dettagli

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.999    0.001    0.997      0.999    0.998      0.997    0.999     0.996     normal.
               0.000    0.000    0.000      0.000    0.000      -0.000   0.498     0.000     buffer_overflow.
               1.000    0.000    0.999      1.000    1.000      0.999    1.000     1.000     neptune.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     smurf.
               0.000    0.000    ?          0.000    ?          ?        0.916     0.001     guess_passwd.
               0.962    0.000    0.962      0.962    0.962      0.962    0.980     0.919     pod.
               1.000    0.000    0.990      1.000    0.995      0.995    1.000     1.000     teardrop.
               0.962    0.000    1.000      0.962    0.980      0.981    0.986     0.970     portsweep.
               0.992    0.000    0.925      0.992    0.957      0.958    0.996     0.870     ipsweep.
               0.000    0.000    ?          0.000    ?          ?        0.750     0.500     land.
               0.986    0.000    0.995      0.986    0.991      0.991    0.995     0.991     back.
               0.000    0.000    ?          0.000    ?          ?        0.500     0.000     imap.
               0.962    0.000    0.981      0.962    0.971      0.971    1.000     0.950     satan.
               0.435    0.000    0.909      0.435    0.588      0.629    0.956     0.422     nmap.
               0.000    0.000    0.000      0.000    0.000      -0.000   1.000     0.216     warezmaster.
               0.873    0.000    0.947      0.873    0.908      0.909    0.992     0.822     warezclient.
               0.000    0.000    ?          0.000    ?          ?        0.500     0.000     rootkit.
Weighted Avg.  0.999    0.000    ?          0.999    ?          ?        1.000     0.997
```

- **La TP Rate** (percentuale di true positive) risulta essere molto alto, tendente al 100%
- **La FP Rate** (percentuale di false positive) risulta essere nulla 0%

- **La FN Rate** (percentuale di falsi negativi) risulta essere nulla 0% (con un grado di confidenza di 0.01, ottima per modellare sistemi critici come in questo caso)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.01
Analysing:  False_negative_rate
Datasets:   1
Resultsets: 1
Confidence: 0.01 (two tailed)
Sorted by:  -
Date:       11/6/20, 4:08 PM


Dataset                   (1) trees.J48
----------------------------------------
'kdd_10_percent-weka.filt(100)   0.00 |
----------------------------------------
(v/ /*)                                |


Key:
(1) trees.J48 '-C 0.25 -M 2' -217733168393644444
```

test effettuato nella modalità Experiment

Di conseguenza si potrebbe concludere che l'algoritmo J48, su questo dataset, funziona in maniera ottima.

Osservazioni personali

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.999	0.001	0.997	0.999	0.998	0.997	0.999	0.996	normal.
	0.000	0.000	0.000	0.000	0.000	-0.000	0.498	0.000	buffer_overflow.
	1.000	0.000	0.999	1.000	1.000	0.999	1.000	1.000	neptune.
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	smurf.
	0.000	0.000	?	0.000	?	?	0.916	0.001	guess_passwd.
	0.962	0.000	0.962	0.962	0.962	0.962	0.980	0.919	pod.
	1.000	0.000	0.990	1.000	0.995	0.995	1.000	1.000	teardrop.
	0.962	0.000	1.000	0.962	0.980	0.981	0.986	0.970	portsweep.
	0.992	0.000	0.925	0.992	0.957	0.958	0.996	0.870	ipsweep.
	0.000	0.000	?	0.000	?	?	0.750	0.500	land.
	0.986	0.000	0.995	0.986	0.991	0.991	0.995	0.991	back.
	0.000	0.000	?	0.000	?	?	0.500	0.000	imap.
	0.962	0.000	0.981	0.962	0.971	0.971	1.000	0.950	satan.
	0.435	0.000	0.909	0.435	0.588	0.629	0.956	0.422	nmap.
	0.000	0.000	0.000	0.000	0.000	-0.000	1.000	0.216	warezmaster.
	0.873	0.000	0.947	0.873	0.908	0.909	0.992	0.822	warezclient.
	0.000	0.000	?	0.000	?	?	0.500	0.000	rootkit.
Weighted Avg.	0.999	0.000	?	0.999	?	?	1.000	0.997	

```
=== Confusion Matrix ===

    a     b     c     d     e     f     g     h     i     j     k     l     m     n     o     p     q   <-- classified as
 9713     1     2     2     0     0     0     0     0     0     1     0     2     0     1     5     0 |     a = normal.
    2     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0 |     b = buffer_overflow.
    1     0 10719     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     c = neptune.
    0     0     0 28078     0     1     0     0     0     0     0     0     0     0     0     0     0 |     d = smurf.
    5     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     e = guess_passwd.
    1     0     0     0     0    25     0     0     0     0     0     0     0     0     0     0     0 |     f = pod.
    0     0     0     0     0     0    97     0     0     0     0     0     0     0     0     0     0 |     g = teardrop.
    2     0     0     0     0     0     0   100     1     0     0     0     1     0     0     0     0 |     h = portsweep.
    1     0     0     0     0     0     0     0   123     0     0     0     0     0     0     0     0 |     i = ipsweep.
    1     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     j = land.
    3     0     0     0     0     0     0     0     0     0   217     0     0     0     0     0     0 |     k = back.
    0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0 |     l = imap.
    0     0     6     0     0     0     0     0     0     0     0     0   152     0     0     0     0 |     m = satan.
    4     0     0     0     0     0     0     0     9     0     0     0     0    10     0     0     0 |     n = nmap.
    1     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     o = warezmaster.
   13     0     0     0     0     0     0     0     0     0     0     0     0     0     0    89     0 |     p = warezclient.
    0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0 |     q = rootkit.
```

1) Le classi cerchiate e sottolineate in rosso, presentano a (mio avviso) dei problemi:

- **accuratezza nulla:** il classificatore le ha classificate tutte erroneamente
- **numero delle istanze pressochè nullo:** molto probabilmente è questa la causa dell' accuratezza così bassa
- **costituiscono circa il 35% dell' insieme delle classi**

Domanda: *in linea di massima questo vorrebbe dire che il nostro classificatore (addestrato in questa maniera) sarebbe vulnerabile al 35% degli attacchi conosciuti e studiati (oltre forse a quelli non)?*

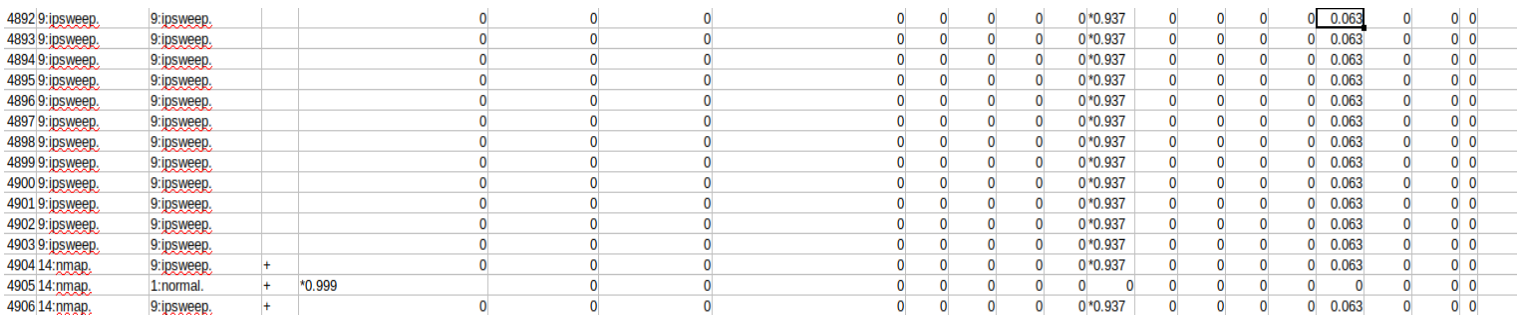
Domanda: *sarebbe possibile (e sensato) aggiungere solamente istanze di queste classi al training set, in maniera tale da allenare maggiormente il modello su queste classi?*

Domanda: *Se la cross-validation divide il training set del 90% e usa un 10% per il test, avendo una sola istanza (come imap), il classificatore riuscirebbe a studiarla ma non a testarla (o vicersa), perchè quell'unica istanza farebbe parte unicamente solo del training set o del test set, corretto? Quindi otterrei sempre 0% di precisione*

Curiosità personale: *la versione nuova del dataset KDD99, dato che classifica solo in* ***{normale, anomalia}*** *potrebbe essere effettivamente migliore? Dato che tuttavia, anche istanze malclassificate (come nel caso delle classi cerchiate in rosso) verrebbero lo stesso classificate come anormali?*

2) Su un nmap, cerchiata in verde, invece l' accuratezza si dimostra essere medio-bassa:
Domanda: *aggiungendo altre istanze (o lavorando sugli attributi) si potrebbe aumentare?*

3) Ipsweep e nmap vengono scambiati con molta facilità, è possibile notarlo anche dalle distribuzioni delle probabilità sul file CSV.



4892	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4893	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4894	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4895	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4896	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4897	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4898	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4899	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4900	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4901	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4902	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4903	9:ipsweep.	9:ipsweep.			0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4904	14:nmap.	9:ipsweep.	+		0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0
4905	14:nmap.	1:normal.	+	*0.999		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4906	14:nmap.	9:ipsweep.	+		0	0	0	0	0	0	0	0	*0.937	0	0	0	0	0.063	0	0	0

4) Le istanze **buffer_overflow** hanno lo 0% di probabilità di essere classificate come tali (file CSV), mentre vengono riconosciuti come normali o warezmaster

5) Le istanze **guess_passwd** hanno lo 0.1% di probabilità di essere classificate come tali, mentre il 99% come normali

6) Le istanze **land** hanno una probabilità inferiore al 33% di essere classificate come tali:

- istanza 1: 0%
- istanza 2: 33%

7) L' istanza **imap** ha una probabilità dello 0% di essere classificata come tale *(questo può essere collegato alla domanda fatta prima sulla cross-validation?)*

8) Le istanze **warezmaster** presentano una probabilità media di essere classificate come tali circa del 25%

- istanza 1: 1.4%
- istanza 2: 50%

9) L' istanza **rootkit**, come imap, presenta una probabilità dello 0% di essere classificata come tale

Studio del dataset con Naive Bayes (con impostazioni di deafult)

Accuratezza

```
=== Summary ===

Correctly Classified Instances       46026               93.1814 %
Incorrectly Classified Instances      3368                6.8186 %
Kappa statistic                          0.8871
Mean absolute error                      0.0078
Root mean squared error                  0.0823
Relative absolute error                 11.2486 %
Root relative squared error             44.1305 %
Total Number of Instances            49394
```

L' accuratezza data risulta essere molto buona (circa 93%), tuttavia rimane peggiore rispetto a quella dell' algoritmo J48 (circa il 99.86%).

La minore precisione e affidabilità del classificatore, rispetto all' algoritmo J48, vengono confermati anche dagli altri valori, descritti successivamente alla percentuale di accuratezza, ed inferiori rispetto a quelli dati dall' algoritmo J48.

Tempo creazione modello

Un vantaggio dell' algoritmo lo si può osservare tuttavia nel tempo di creazione del modello:

```
Time taken to build model: 0.32 seconds
```

(Naive bayes)

contro:

```
Time taken to build model: 1.01 seconds
```

(J48)

Dettagli

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.001    0.996      0.667    0.799      0.783    0.998     0.993     normal.
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.699     0.033     buffer_overflow.
                 0.999    0.000    0.999      0.999    0.999      0.999    1.000     1.000     neptune.
                 0.999    0.000    1.000      0.999    1.000      0.999    1.000     1.000     smurf.
                 0.800    0.033    0.002      0.800    0.005      0.043    0.993     0.426     guess_passwd.
                 1.000    0.001    0.433      1.000    0.605      0.658    1.000     0.921     pod.
                 0.979    0.001    0.679      0.979    0.802      0.815    0.999     0.654     teardrop.
                 0.942    0.003    0.405      0.942    0.566      0.617    0.999     0.598     portsweep.
                 0.935    0.001    0.648      0.935    0.766      0.778    0.996     0.815     ipsweep.
                 0.000    0.006    0.000      0.000    0.000      -0.001   0.923     0.000     land.
                 0.950    0.000    1.000      0.950    0.974      0.975    1.000     0.996     back.
                 0.000    0.000    ?          0.000    ?          ?        0.872     0.000     imap.
                 0.854    0.001    0.675      0.854    0.754      0.759    0.994     0.833     satan.
                 0.565    0.010    0.026      0.565    0.050      0.120    0.985     0.074     nmap.
                 0.500    0.000    0.045      0.500    0.083      0.151    0.994     0.252     warezmaster.
                 0.716    0.008    0.159      0.716    0.260      0.335    0.996     0.348     warezclient.
                 0.000    0.003    0.000      0.000    0.000      -0.000   0.911     0.000     rootkit.
Weighted Avg.    0.932    0.000    ?          0.932    ?          ?        1.000     0.994
```

- **La TP Rate** (percentuale di true positive) risulta essere alta, tendente al 93%, tuttavia inferiore a J48 (tendente al 100%)
- **La FP Rate** (percentuale di false positive) risulta essere nulla 0% (come J48)
- **La FN Rate** (percentuale di falsi negativi) risulta essere del 33% (con un grado di confidenza di 0.01, ottima per modellare sistemi critici come in questo caso), risultato non buono a mio modesto avviso, dato che mediamente 1 attacco su 3 potrebbe non essere riconosciuto come tale

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.01 -result-matrix "weka.experiment.ResultMat
Analysing:  False_negative_rate
Datasets:   1
Resultsets: 1
Confidence: 0.01 (two tailed)
Sorted by:  -
Date:       11/7/20, 10:30 AM


Dataset                   (1) bayes.Nai
----------------------------------------
'kdd_10_percent-weka.filt(100)   0.33 |
----------------------------------------
(v/ /*)                               |


Key:
(1) bayes.NaiveBayes '' 5995231201785697655
```

Osservazioni personali

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.001    0.996      0.667    0.799      0.783    0.998     0.993     normal.
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.699     0.033     buffer_overflow.
                 0.999    0.000    0.999      0.999    0.999      0.999    1.000     1.000     neptune.
                 0.999    0.000    1.000      0.999    1.000      0.999    1.000     1.000     smurf.
                 0.800    0.033    0.002      0.800    0.005      0.043    0.993     0.426     guess_passwd.
                 1.000    0.001    0.433      1.000    0.605      0.658    1.000     0.921     pod.
                 0.979    0.001    0.679      0.979    0.802      0.815    0.999     0.654     teardrop.
                 0.942    0.003    0.405      0.942    0.566      0.617    0.999     0.598     portsweep.
                 0.935    0.001    0.648      0.935    0.766      0.778    0.996     0.815     ipsweep.
                 0.000    0.006    0.000      0.000    0.000      -0.001   0.923     0.000     land.
                 0.950    0.000    1.000      0.950    0.974      0.975    1.000     0.996     back.
                 0.000    0.000    ?          0.000    ?          ?        0.872     0.000     imap.
                 0.854    0.001    0.675      0.854    0.754      0.759    0.994     0.833     satan.
                 0.565    0.010    0.026      0.565    0.050      0.120    0.985     0.074     nmap.
                 0.500    0.000    0.045      0.500    0.083      0.151    0.994     0.252     warezmaster.
                 0.716    0.008    0.159      0.716    0.260      0.335    0.996     0.348     warezclient.
                 0.000    0.003    0.000      0.000    0.000      -0.000   0.911     0.000     rootkit.
Weighted Avg.    0.932    0.000    ?          0.932    ?          ?        1.000     0.994
```

```
=== Confusion Matrix ===

    a     b     c     d     e     f     g     h     i     j     k     l     m     n     o     p     q   <-- classified as
 6492     8     0     1  1619    27    38   124    54   314     0     0    51   475    20   385   119 |     a = normal.
    1     0     0     0     0     0     0     1     0     0     0     0     0     0     1     0     0 |     b = buffer_overflow.
    0     0 10705     0     2     0     0     2     0     2     0     0     9     0     0     0     0 |     c = neptune.
    9     0     0 28059     0     4     7     0     0     0     0     0     0     0     0     0     0 |     d = smurf.
    0     0     0     0     4     0     0     0     0     0     0     0     0     0     0     1     0 |     e = guess_passwd.
    0     0     0     0     0    26     0     0     0     0     0     0     0     0     0     0     0 |     f = pod.
    0     0     0     0     0     2    95     0     0     0     0     0     0     0     0     0     0 |     g = teardrop.
    0     0     0     0     0     0     0    98     0     0     0     0     5     1     0     0     0 |     h = portsweep.
    0     0     0     0     0     0     0     7   116     1     0     0     0     0     0     0     0 |     i = ipsweep.
    0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     1 |     j = land.
   11     0     0     0     0     0     0     0     0     0   209     0     0     0     0     0     0 |     k = back.
    0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0 |     l = imap.
    4     1     6     0     0     1     0    10     0     0     0     0   135     1     0     0     0 |     m = satan.
    1     0     0     0     0     0     0     0     9     0     0     0     0    13     0     0     0 |     n = nmap.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0 |     o = warezmaster.
    0     0     0     0     0     0     0     0     0     0     0     0     0     2     0    73    27 |     p = warezclient.
    0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0 |     q = rootkit.
```

1) **normal:** buoni risultati, simili a quelli dati da J48, anche se leggermente inferiori

2) **buffer_overflow:** la precisione ottenuta è nulla come in J48, tuttavia, in questo caso le istanze di questa classe presentano una probabilità media più alta di essere riconosciute come tali (file CSV):
    1. istanza 4941 = 0%
    2. istanza 9881 = 9%
    3. istanza 49395 = 5.8%

   Questo è evidenziato anche dal fatto anche che la PRC area è maggiore rispetto a quella di J48, dove il valore era di 0.

3) **guess_passwd:** in questo caso, il Naive bayes, si è dimostrato nettamente migliore al J48, dato che è riusciuto a classificare correttamente 4 istanze su 5 (mentre J48 nessuna). Le istanze guess_passwd, presentano una probabilità media dell' 86% di essere classificate come tali, mentre in J48 questa probabilità era dello 0%

4) **warezmaster**: anche in questo caso, il Naive bayes, si è dimostrato migliore dell' algoritmo J48, riuscendo a classificare correttamente 1 istanza (su 2), mentre J48 non è stato in grado

5) **imap** e **rootkit**: come J48, il Naive bayes li ha classificati erroneamente, anche se c'è da ricordare che queste sono le classi che presentano un' unica istanza

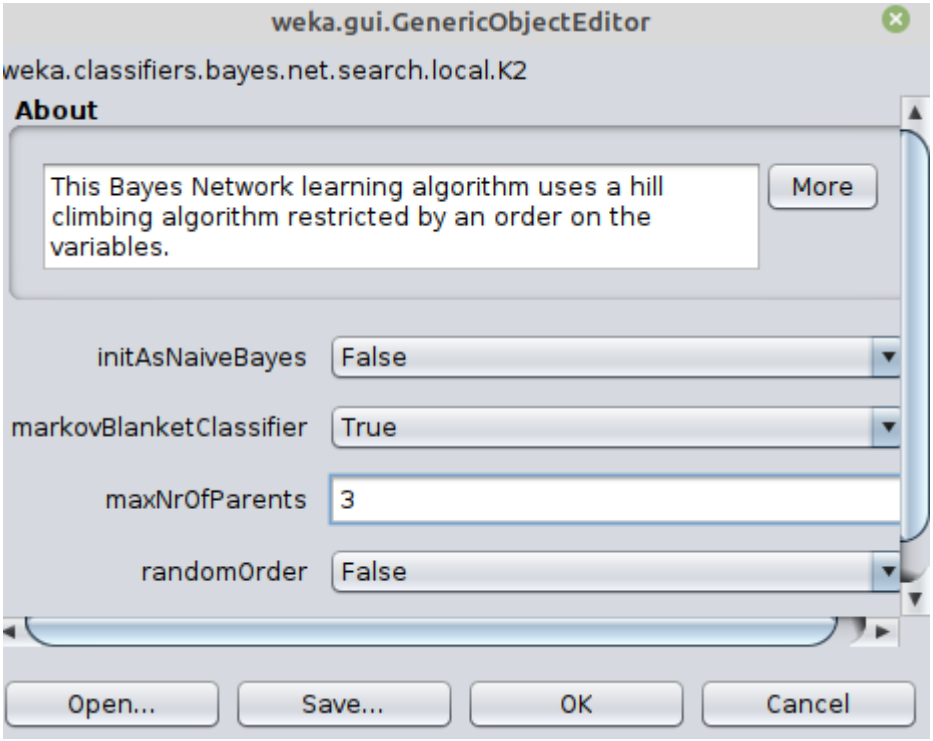
6) tutte le altre classi non le ho citate perchè hanno ottenuto un comportamento più o meno simile (o peggiore) rispetto a J48

Domanda: *il fatto che l' algoritmo NB sia riuscito a classificare (nonostante avendo un accuratezza inferiore) correttamente le classi con un minor numero di istanze (come buffer_overflow e guess_passwd), può essere sintomo di overfitting nel decision tree creato da J48?*

Curiosità personale: *in questo dataset risulta essere difficile (a parer mio), poter calcolare la FN rate (la percentuale di falsi negativi), perchè anche se il nostro classificare sbagliasse a classificarle come un attacco di tipo B, potrebbe classificarli con un attacco di tipo C, ma in ogni caso verrebbe segnalato dal sistema. La versione migliorata del dataset KDD, avendo solo due classi invece, potrebbe fornire una visione molto più chiara del FN rate?*

Studio del dataset con BayesNet (con impostazioni personalizzate)

Impostazioni



Accuratezza

```
Correctly Classified Instances       49330               99.8704 %
Incorrectly Classified Instances        64                0.1296 %
Kappa statistic                          0.9978
Mean absolute error                      0.0002
Root mean squared error                  0.0115
Relative absolute error                  0.2759 %
Root relative squared error              6.1539 %
Total Number of Instances            49394
```

L' accuratezza data risulta essere estremamente ottima, circa il 99.87%, più alta anche dell' algoritmo J48.
La percentuale di istanze classificate erroneamente, di conseguenza, anche è più bassa.
La precisione e l'affidabilità del classificatore vengono confermati anche dagli altri valori, descritti successivamente alla percentuale di accuratezza, e leggermenti migliori di quelli presentati da J48.

Tempo creazione modello

```
Time taken to build model: 1.85 seconds
```

Maggiore rispetto a tutti i suoi predecessori

Dettagli

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.998	0.001	0.998	0.998	0.998	0.998	1.000	1.000	normal.
	0.000	0.000	?	0.000	?	?	0.998	0.052	buffer_overflow.
	1.000	0.000	0.999	1.000	0.999	0.999	1.000	1.000	neptune.
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	smurf.
	0.800	0.000	0.667	0.800	0.727	0.730	1.000	0.808	guess_passwd.
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	pod.
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	teardrop.
	0.971	0.000	0.962	0.971	0.967	0.966	1.000	0.988	portsweep.
	0.960	0.000	0.930	0.960	0.944	0.944	1.000	0.933	ipsweep.
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	land.
	1.000	0.000	0.982	1.000	0.991	0.991	1.000	1.000	back.
	0.000	0.000	?	0.000	?	?	1.000	0.059	imap.
	0.911	0.000	0.980	0.911	0.944	0.945	1.000	0.987	satan.
	0.435	0.000	0.833	0.435	0.571	0.602	0.999	0.538	nmap.
	0.000	0.000	?	0.000	?	?	1.000	0.076	warezmaster.
	0.951	0.000	0.951	0.951	0.951	0.951	1.000	0.975	warezclient.
	0.000	0.000	?	0.000	?	?	0.998	0.008	rootkit.
Weighted Avg.	0.999	0.000	?	0.999	?	?	1.000	0.999	

- **La TP Rate** (percentuale di true positive) risulta essere ottima, tendente al 100% come J48
- **La FP Rate** (percentuale di false positive) risulta essere nulla 0% (come J48)
- **La FN Rate** (percentuale di falsi negativi) risulta essere dello 0% (con un grado di confidenza di 0.01)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.01 -result-matrix "weka.experiment.ResultMat
Analysing:  False_negative_rate
Datasets:   1
Resultsets: 1
Confidence: 0.01 (two tailed)
Sorted by:  -
Date:       11/7/20, 12:31 PM


Dataset                   (1) bayes.Bay
----------------------------------------
'kdd_10_percent-weka.filt(100)   0.00 |
----------------------------------------
(v/ /*)                               |


Key:
(1) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 3 -N -mbc -S BAYES -E bayes.net.estimate.SimpleEstimator -- -
```

Osservazioni personali

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.998     0.001     0.998      0.998     0.998      0.998     1.000     1.000     normal.
               0.000     0.000     ?          0.000     ?          ?         0.998     0.052     buffer_overflow.
               1.000     0.000     0.999      1.000     0.999      0.999     1.000     1.000     neptune.
               1.000     0.000     1.000      1.000     1.000      1.000     1.000     1.000     smurf.
               0.800     0.000     0.667      0.800     0.727      0.730     1.000     0.808     guess_passwd.
               1.000     0.000     1.000      1.000     1.000      1.000     1.000     1.000     pod.
               1.000     0.000     1.000      1.000     1.000      1.000     1.000     1.000     teardrop.
               0.971     0.000     0.962      0.971     0.967      0.966     1.000     0.988     portsweep.
               0.960     0.000     0.930      0.960     0.944      0.944     1.000     0.933     ipsweep.
               1.000     0.000     1.000      1.000     1.000      1.000     1.000     1.000     land.
               1.000     0.000     0.982      1.000     0.991      0.991     1.000     1.000     back.
               0.000     0.000     ?          0.000     ?          ?         1.000     0.059     imap.
               0.911     0.000     0.980      0.911     0.944      0.945     1.000     0.987     satan.
               0.435     0.000     0.833      0.435     0.571      0.602     0.999     0.538     nmap.
               0.000     0.000     ?          0.000     ?          ?         1.000     0.076     warezmaster.
               0.951     0.000     0.951      0.951     0.951      0.951     1.000     0.975     warezclient.
               0.000     0.000     ?          0.000     ?          ?         0.998     0.008     rootkit.
Weighted Avg.  0.999     0.000     ?          0.999     ?          ?         1.000     0.999

=== Confusion Matrix ===

    a     b     c     d     e     f     g     h     i     j     k     l     m     n     o     p     q   <-- classified as
 9711     0     0     0     2     0     0     4     0     0     4     0     2     1     0     3     0 |     a = normal.
    1     0     1     0     0     0     0     0     0     0     0     0     0     0     0     1     0 |     b = buffer_overflow.
    0     0 10720     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     c = neptune.
    0     0     0 28079     0     0     0     0     0     0     0     0     0     0     0     0     0 |     d = smurf.
    1     0     0     0     4     0     0     0     0     0     0     0     0     0     0     0     0 |     e = guess_passwd.
    0     0     0     0     0    26     0     0     0     0     0     0     0     0     0     0     0 |     f = pod.
    0     0     0     0     0     0    97     0     0     0     0     0     0     0     0     0     0 |     g = teardrop.
    0     0     2     0     0     0     0   101     0     0     0     0     1     0     0     0     0 |     h = portsweep.
    1     0     4     0     0     0     0     0   119     0     0     0     0     0     0     0     0 |     i = ipsweep.
    0     0     0     0     0     0     0     0     0     2     0     0     0     0     0     0     0 |     j = land.
    0     0     0     0     0     0     0     0     0     0   220     0     0     0     0     0     0 |     k = back.
    0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0 |     l = imap.
    6     0     8     0     0     0     0     0     0     0     0     0   144     0     0     0     0 |     m = satan.
    4     0     0     0     0     0     0     0     9     0     0     0     0    10     0     0     0 |     n = nmap.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0 |     o = warezmaster.
    5     0     0     0     0     0     0     0     0     0     0     0     0     0     0    97     0 |     p = warezclient.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     q = rootkit.
```

1) normal: simile a J48

2) buffer_overflow: la precisione ottenuta è nulla come in J48. La probabilità di essere riconosciute come tali (file CSV):

4. istanza 4941 = 0.7%
5. istanza 9881 = 0%
6. istanza 49395 = 0%

Questo evidenzia il fatto che Naive bayes, in termini probabilistici, è migliore

3) neptune, smurf,pod,teardrop,back: precisione del 100%

4) guess_passwd: come Naive bayes, 4 istanze classificate correttamente su 5

5) land: primo classificatore a classificare non solo correttamente una istanza, ma tutte quante. Precisione del 100%

6) warezmaster: precisione dello 0% come J48. Naive bayes ancora il migliore con una precisione del 50%

7) warezclient: precisione del 95%, la migliore finora

8) imap, rootkit: come per gli altri classificatori, precisione dello 0%.

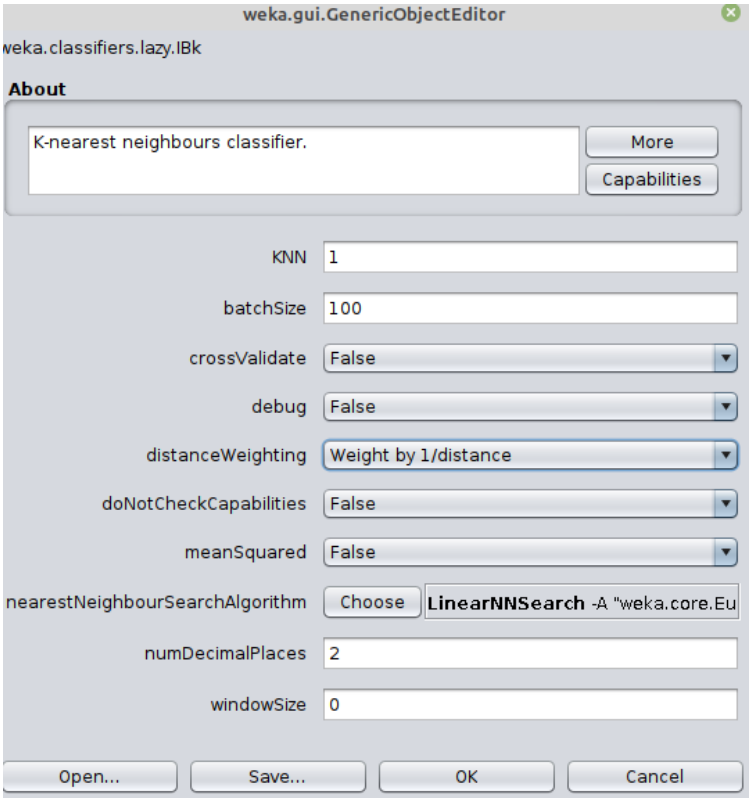
In definitiva si potrebbe dire che la rete bayesiana abbia prodotto, per ora, i risultati mediamente migliori; essendo più o meno preciso come J48 e avendo preso in parte alcuni vantaggi del Naive bayes non avuti da J48.

Studio del dataset con IBk (con impostazioni personalizzate)

Implementazione dell' algoritmo kNN in WEKA.

K=1

Impostazioni



Di norma l'algoritmo presuppone che gli attributi siano equi-pesati (stessa importanza). Tuttavia in questo caso gli attribuiamo un peso inversamente proporzionale alla distanza (più è vicino, quindi meno è distante, maggiore è il suo peso)

Accuratezza

```
Correctly Classified Instances       49336               99.8826 %
Incorrectly Classified Instances        58                0.1174 %
Kappa statistic                          0.998
Mean absolute error                      0.0001
Root mean squared error                  0.0118
Relative absolute error                  0.2117 %
Root relative squared error              6.3202 %
Total Number of Instances            49394
```

L' accuratezza data risulta essere estremamente ottima, poco più del 99.88%, più alta sia della rete bayesiana, sia dell' algoritmo J48.

La percentuale di istanze classificate erroneamente, di conseguenza, anche è più bassa.
La precisione e l'affidabilità del classificatore vengono confermati anche dagli altri valori, descritti successivamente alla percentuale di accuratezza.

Tempo creazione modello

```
Time taken to build model: 0.01 seconds
```

Essendo un algoritmo *lazy*, l' algoritmo non fa nulla finchè non viene chiesto di eseguire una query su di esso. Di conseguenza la creazione del modello è unicamente una raccolta dati, questo spiega il tempo di creazione del modello.
Tuttavia, la sua completa esecuzione in Cross-Validation (10), porta via mediamente piu o meno **10 minuti** di tempo (sul mio computer), e ad ora è l'algoritmo più lento a restituire risultati.

Dettagli

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.999	0.001	0.998	0.999	0.998	0.998	1.000	1.000	normal.
	0.000	0.000	0.000	0.000	0.000	-0.000	0.866	0.181	buffer_overflow.
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	neptune.
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	smurf.
	0.800	0.000	0.667	0.800	0.727	0.730	1.000	0.819	guess_passwd.
	1.000	0.000	0.963	1.000	0.981	0.981	1.000	1.000	pod.
	0.979	0.000	1.000	0.979	0.990	0.990	1.000	0.989	teardrop.
	0.971	0.000	1.000	0.971	0.985	0.985	1.000	0.991	portsweep.
	0.952	0.000	0.922	0.952	0.937	0.936	0.990	0.910	ipsweep.
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	land.
	0.982	0.000	0.991	0.982	0.986	0.986	1.000	0.985	back.
	0.000	0.000	?	0.000	?	?	1.000	0.500	imap.
	0.975	0.000	0.994	0.975	0.984	0.984	1.000	0.993	satan.
	0.565	0.000	0.650	0.565	0.605	0.606	0.915	0.546	nmap.
	0.500	0.000	1.000	0.500	0.667	0.707	0.999	0.508	warezmaster.
	0.941	0.000	0.906	0.941	0.923	0.923	0.995	0.944	warezclient.
	0.000	0.000	0.000	0.000	0.000	-0.000	0.996	0.005	rootkit.
Weighted Avg.	0.999	0.000	?	0.999	?	?	1.000	0.999	

- **La TP Rate** (percentuale di true positive) risulta essere ottima, tendente al 100%
- **La FP Rate** (percentuale di false positive) risulta essere nulla 0%
- **La FN Rate** (percentuale di falsi negativi) risulta essere dello 0% (con un grado di confidenza di 0.01)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3,4,5 -D 1 -R 2 -S 0.01 -resul
Analysing:  False_negative_rate
Datasets:   1
Resultsets: 1
Confidence: 0.01 (two tailed)
Sorted by:  -
Date:       11/9/20, 3:05 PM


Dataset                   (1) lazy.IBk
----------------------------------------
'kdd_10_percent-weka.filt  (1)   0.00 |
----------------------------------------
(v/ /*)                               |


Key:
(1) lazy.IBk '-K 1 -W 0 -I -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka
```

Osservazioni personali

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.999    0.001    0.998      0.999    0.998      0.998    1.000     1.000     normal.
               0.000    0.000    0.000      0.000    0.000      -0.000   0.866     0.181     buffer_overflow.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     neptune.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     smurf.
               0.800    0.000    0.667      0.800    0.727      0.730    1.000     0.819     guess_passwd.
               1.000    0.000    0.963      1.000    0.981      0.981    1.000     1.000     pod.
               0.979    0.000    1.000      0.979    0.990      0.990    1.000     0.989     teardrop.
               0.971    0.000    1.000      0.971    0.985      0.985    1.000     0.991     portsweep.
               0.952    0.000    0.922      0.952    0.937      0.936    0.990     0.910     ipsweep.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     land.
               0.982    0.000    0.991      0.982    0.986      0.986    1.000     0.985     back.
               0.000    0.000    ?          0.000    ?          ?        1.000     0.500     imap.
               0.975    0.000    0.994      0.975    0.984      0.984    1.000     0.993     satan.
               0.565    0.000    0.650      0.565    0.605      0.606    0.915     0.546     nmap.
               0.500    0.000    1.000      0.500    0.667      0.707    0.999     0.508     warezmaster.
               0.941    0.000    0.906      0.941    0.923      0.923    0.995     0.944     warezclient.
               0.000    0.000    0.000      0.000    0.000      -0.000   0.996     0.005     rootkit.
Weighted Avg.  0.999    0.000    ?          0.999    ?          ?        1.000     0.999

=== Confusion Matrix ===

    a     b     c     d     e     f     g     h     i     j     k     l     m     n     o     p     q   <-- classified as
 9714     0     0     0     1     0     0     0     0     0     2     0     0     0     0     9     1 |     a = normal.
    0     0     1     0     1     0     0     0     0     0     0     0     0     0     0     1     0 |     b = buffer_overflow.
    1     0 10719     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     c = neptune.
    1     0     0 28077     0     1     0     0     0     0     0     0     0     0     0     0     0 |     d = smurf.
    0     1     0     0     4     0     0     0     0     0     0     0     0     0     0     0     0 |     e = guess_passwd.
    0     0     0     0     0    26     0     0     0     0     0     0     0     0     0     0     0 |     f = pod.
    2     0     0     0     0     0    95     0     0     0     0     0     0     0     0     0     0 |     g = teardrop.
    1     0     1     0     0     0     0   101     1     0     0     0     0     0     0     0     0 |     h = portsweep.
    1     0     0     0     0     0     0     0   118     0     0     0     0     5     0     0     0 |     i = ipsweep.
    0     0     0     0     0     0     0     0     0     2     0     0     0     0     0     0     0 |     j = land.
    4     0     0     0     0     0     0     0     0     0   216     0     0     0     0     0     0 |     k = back.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     l = imap.
    2     0     0     0     0     0     0     0     0     0     0     0   154     2     0     0     0 |     m = satan.
    0     0     0     0     0     0     0     0     9     0     0     0     1    13     0     0     0 |     n = nmap.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0 |     o = warezmaster.
    6     0     0     0     0     0     0     0     0     0     0     0     0     0     0    96     0 |     p = warezclient.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     q = rootkit.
```

1) normal: simile a J48

2) **buffer_overflow:** la precisione ottenuta è nulla come in J48. La probabilità di essere riconosciute come tali (file CSV):

1. istanza 4941 = 0%
2. istanza 9881 = 0%
3. istanza 49395 = 0%

3) **neptune, smurf, pod, land:** precisione del 100%

4) **guess_passwd:** come Naive bayes, 4 istanze classificate correttamente su 5

5) **land:** secondo classificatore con precisione del 100%

6) **warezmaster:** precisione del 50% come Naive bayes

7) **imap, rootkit:** la situazione rimane la stessa

K=3

```
=== Summary ===

Correctly Classified Instances       49330               99.8704 %
Incorrectly Classified Instances        64                0.1296 %
Kappa statistic                          0.9978
Mean absolute error                      0.0002
Root mean squared error                  0.0112
Relative absolute error                  0.2385 %
Root relative squared error              6.007  %
Total Number of Instances            49394

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.001    0.998      0.998    0.998      0.997    1.000     1.000     normal.
                 0.000    0.000    ?          0.000    ?          ?        0.866     0.127     buffer_overflow.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     neptune.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     smurf.
                 0.800    0.000    0.571      0.800    0.667      0.676    1.000     0.819     guess_passwd.
                 1.000    0.000    0.963      1.000    0.981      0.981    1.000     1.000     pod.
                 0.979    0.000    1.000      0.979    0.990      0.990    1.000     0.990     teardrop.
                 0.971    0.000    0.971      0.971    0.971      0.971    1.000     0.990     portsweep.
                 0.976    0.000    0.924      0.976    0.949      0.949    1.000     0.948     ipsweep.
                 0.000    0.000    ?          0.000    ?          ?        1.000     1.000     land.
                 0.977    0.000    0.991      0.977    0.984      0.984    1.000     0.988     back.
                 0.000    0.000    ?          0.000    ?          ?        1.000     0.333     imap.
                 0.962    0.000    0.993      0.962    0.977      0.978    1.000     0.997     satan.
                 0.565    0.000    0.765      0.565    0.650      0.657    0.938     0.659     nmap.
                 0.000    0.000    ?          0.000    ?          ?        0.999     0.507     warezmaster.
                 0.941    0.000    0.881      0.941    0.910      0.910    0.996     0.940     warezclient.
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.997     0.007     rootkit.
Weighted Avg.    0.999    0.000    ?          0.999    ?          ?        1.000     0.999

=== Confusion Matrix ===

    a     b     c     d     e     f     g     h     i     j     k     l     m     n     o     p     q   <-- classified as
 9711     0     0     0     1     0     0     0     0     0     2     0     0     0     0    12     1 |     a = normal.
    1     0     1     0     0     0     0     0     0     0     0     0     0     0     0     1     0 |     b = buffer_overflow.
    0     0 10719     0     1     0     0     0     0     0     0     0     0     0     0     0     0 |     c = neptune.
    1     0     0 28077     0     1     0     0     0     0     0     0     0     0     0     0     0 |     d = smurf.
    1     0     0     0     4     0     0     0     0     0     0     0     0     0     0     0     0 |     e = guess_passwd.
    0     0     0     0     0    26     0     0     0     0     0     0     0     0     0     0     0 |     f = pod.
    2     0     0     0     0     0    95     0     0     0     0     0     0     0     0     0     0 |     g = teardrop.
    0     0     1     0     1     0     0   101     1     0     0     0     0     0     0     0     0 |     h = portsweep.
    1     0     0     0     0     0     0     0   121     0     0     0     0     2     0     0     0 |     i = ipsweep.
    2     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     j = land.
    5     0     0     0     0     0     0     0     0     0   215     0     0     0     0     0     0 |     k = back.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     l = imap.
    1     0     0     0     0     0     0     3     0     0     0     0   152     2     0     0     0 |     m = satan.
    0     0     0     0     0     0     0     0     9     0     0     0     1    13     0     0     0 |     n = nmap.
    2     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     o = warezmaster.
    6     0     0     0     0     0     0     0     0     0     0     0     0     0     0    96     0 |     p = warezclient.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     q = rootkit.
```

K=5

```
=== Summary ===

Correctly Classified Instances       49333               99.8765 %
Incorrectly Classified Instances        61                0.1235 %
Kappa statistic                          0.9979
Mean absolute error                      0.0002
Root mean squared error                  0.0114
Relative absolute error                  0.2689 %
Root relative squared error              6.1384 %
Total Number of Instances            49394

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.999    0.001    0.997      0.999    0.998      0.998    1.000     1.000     normal.
               0.000    0.000    ?          0.000    ?          ?        0.867     0.181     buffer_overflow.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     neptune.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     smurf.
               0.800    0.000    0.500      0.800    0.615      0.632    1.000     0.818     guess_passwd.
               1.000    0.000    0.963      1.000    0.981      0.981    1.000     1.000     pod.
               0.979    0.000    1.000      0.979    0.990      0.990    1.000     0.989     teardrop.
               0.971    0.000    0.962      0.971    0.967      0.966    1.000     0.990     portsweep.
               0.992    0.000    0.925      0.992    0.957      0.958    1.000     0.952     ipsweep.
               0.000    0.000    ?          0.000    ?          ?        1.000     1.000     land.
               0.977    0.000    0.991      0.977    0.984      0.984    1.000     0.988     back.
               0.000    0.000    ?          0.000    ?          ?        1.000     0.500     imap.
               0.949    0.000    0.993      0.949    0.971      0.971    1.000     0.996     satan.
               0.565    0.000    0.929      0.565    0.703      0.724    0.958     0.664     nmap.
               0.000    0.000    ?          0.000    ?          ?        0.998     0.507     warezmaster.
               0.941    0.000    0.914      0.941    0.928      0.927    0.996     0.938     warezclient.
               0.000    0.000    0.000      0.000    0.000      -0.000   0.997     0.008     rootkit.
Weighted Avg.  0.999    0.000    ?          0.999    ?          ?        1.000     0.999

=== Confusion Matrix ===

    a     b     c     d     e     f     g     h     i     j     k     l     m     n     o     p     q   <-- classified as
 9715     0     0     0     1     0     0     0     0     0     2     0     0     0     0     8     1 |     a = normal.
    1     0     0     0     1     0     0     0     0     0     0     0     0     0     0     1     0 |     b = buffer_overflow.
    0     0 10719     0     1     0     0     0     0     0     0     0     0     0     0     0     0 |     c = neptune.
    2     0     0 28076     0     1     0     0     0     0     0     0     0     0     0     0     0 |     d = smurf.
    1     0     0     0     4     0     0     0     0     0     0     0     0     0     0     0     0 |     e = guess_passwd.
    0     0     0     0     0    26     0     0     0     0     0     0     0     0     0     0     0 |     f = pod.
    2     0     0     0     0     0    95     0     0     0     0     0     0     0     0     0     0 |     g = teardrop.
    0     0     1     0     1     0     0   101     1     0     0     0     0     0     0     0     0 |     h = portsweep.
    1     0     0     0     0     0     0     0   123     0     0     0     0     0     0     0     0 |     i = ipsweep.
    2     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     j = land.
    5     0     0     0     0     0     0     0     0     0   215     0     0     0     0     0     0 |     k = back.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     l = imap.
    3     0     0     0     0     0     0     4     0     0     0     0   150     1     0     0     0 |     m = satan.
    0     0     0     0     0     0     0     0     9     0     0     0     1    13     0     0     0 |     n = nmap.
    2     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     o = warezmaster.
    6     0     0     0     0     0     0     0     0     0     0     0     0     0     0    96     0 |     p = warezclient.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     q = rootkit.
```

Ovviamente aumentando il valore di K, le classi con poche istanze sono state penalizzate.

Ad esempio **land**, avendo solo 2 istanze, con K=1 ha ottenuto una precisione del 100%.
Aumentando K oltre a 2, la probabilità di ottenere un 100% di accuratezza è scesa.

Studio del dataset con Multilayer Perceptron (con impostazioni di default)

Accuratezza

```
=== Summary ===

Correctly Classified Instances       49035               99.2732 %
Incorrectly Classified Instances       359                0.7268 %
Kappa statistic                          0.9877
Mean absolute error                      0.0014
Root mean squared error                  0.0267
Relative absolute error                  1.9552 %
Root relative squared error             14.3021 %
Total Number of Instances            49394
```

L' accuratezza data risulta essere molto buona, tuttavia risulta essere la peggiore dopo quella del Naive Bayes.

Tempistica

```
23:38:27: Command: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
08:12:59: Finished weka.classifiers.functions.MultilayerPerceptron
```

L' addestramento della rete neurale ha richiesto molto tempo in più rispetto alla creazione e alla verifica dei modelli creati dagli altri algoritmi.
Come si può notare dalla figura sovrastante, l' algoritmo ha impiegato circa **8 ore e 30 minuti** per terminare e restituire i risultati.

Dettagli

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.994    0.007    0.974      0.994    0.984      0.980    0.999     0.995     normal.
               0.000    0.000    ?          0.000    ?          ?        0.759     0.000     buffer_overflow.
               1.000    0.000    0.999      1.000    1.000      1.000    1.000     1.000     neptune.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     smurf.
               0.400    0.000    1.000      0.400    0.571      0.632    0.993     0.407     guess_passwd.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     pod.
               0.979    0.000    0.969      0.979    0.974      0.974    1.000     0.999     teardrop.
               0.971    0.000    0.886      0.971    0.927      0.927    1.000     0.992     portsweep.
               0.935    0.001    0.773      0.935    0.847      0.850    0.986     0.783     ipsweep.
               0.000    0.000    ?          0.000    ?          ?        0.875     0.000     land.
               0.068    0.000    0.833      0.068    0.126      0.238    0.983     0.268     back.
               0.000    0.000    ?          0.000    ?          ?        0.598     0.000     imap.
               0.892    0.000    0.993      0.892    0.940      0.941    0.916     0.899     satan.
               0.391    0.000    0.750      0.391    0.514      0.542    0.987     0.495     nmap.
               0.000    0.000    ?          0.000    ?          ?        0.655     0.000     warezmaster.
               0.686    0.001    0.673      0.686    0.680      0.679    0.998     0.759     warezclient.
               0.000    0.000    ?          0.000    ?          ?        0.631     0.000     rootkit.
Weighted Avg.  0.993    0.001    ?          0.993    ?          ?        0.999     0.994
```

- **La TP Rate** (percentuale di true positive) risulta essere molto buona, ma bassa se confrontata con gli altri algoritmi (99.3%)
- **La FP Rate** (percentuale di false positive) risulta essere dello 0.1%

- **La FN Rate** (percentuale di falsi negativi) non è stata calcolata per via del tempo che avrebbe richiesto nuovamente l' algoritmo. Tuttavia, in questo caso, ho ritenuto il dato essere superfluo, dato che, se anche avesse restituito uno 0%, altri algoritmi si sono dimostrati superiori in ogni caso (oltre ad avere anche essi un FN rate dello 0%)

Osservazioni personali

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.994    0.007    0.974      0.994    0.984      0.980    0.999     0.995     normal.
               0.000    0.000    ?          0.000    ?          ?        0.759     0.000     buffer_overflow.
               1.000    0.000    0.999      1.000    1.000      1.000    1.000     1.000     neptune.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     smurf.
               0.400    0.000    1.000      0.400    0.571      0.632    0.993     0.407     guess_passwd.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     pod.
               0.979    0.000    0.969      0.979    0.974      0.974    1.000     0.999     teardrop.
               0.971    0.000    0.886      0.971    0.927      0.927    1.000     0.992     portsweep.
               0.935    0.001    0.773      0.935    0.847      0.850    0.986     0.783     ipsweep.
               0.000    0.000    ?          0.000    ?          ?        0.875     0.000     land.
               0.068    0.000    0.833      0.068    0.126      0.238    0.983     0.268     back.
               0.000    0.000    ?          0.000    ?          ?        0.598     0.000     imap.
               0.892    0.000    0.993      0.892    0.940      0.941    0.916     0.899     satan.
               0.391    0.000    0.750      0.391    0.514      0.542    0.987     0.495     nmap.
               0.000    0.000    ?          0.000    ?          ?        0.655     0.000     warezmaster.
               0.686    0.001    0.673      0.686    0.680      0.679    0.998     0.759     warezclient.
               0.000    0.000    ?          0.000    ?          ?        0.631     0.000     rootkit.
Weighted Avg.  0.993    0.001    ?          0.993    ?          ?        0.999     0.994

=== Confusion Matrix ===

    a     b     c     d     e     f     g     h     i     j     k     l     m     n     o     p     q   <-- classified as
 9664     0     0     1     0     0     0     2    25     0     3     0     0     0     0    32     0 |     a = normal.
    2     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0 |     b = buffer_overflow.
    0     0 10720     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     c = neptune.
    3     0     0 28076     0     0     0     0     0     0     0     0     0     0     0     0     0 |     d = smurf.
    3     0     0     0     2     0     0     0     0     0     0     0     0     0     0     0     0 |     e = guess_passwd.
    0     0     0     0     0    26     0     0     0     0     0     0     0     0     0     0     0 |     f = pod.
    2     0     0     0     0     0    95     0     0     0     0     0     0     0     0     0     0 |     g = teardrop.
    0     0     1     0     0     0     0   101     0     0     0     0     1     1     0     0     0 |     h = portsweep.
    2     0     5     0     0     0     0     1   116     0     0     0     0     0     0     0     0 |     i = ipsweep.
    0     0     1     0     0     0     0     1     0     0     0     0     0     0     0     0     0 |     j = land.
  205     0     0     0     0     0     0     0     0     0    15     0     0     0     0     0     0 |     k = back.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     l = imap.
    6     0     0     0     0     0     3     6     0     0     0     0   141     2     0     0     0 |     m = satan.
    2     0     0     0     0     0     0     3     9     0     0     0     0     9     0     0     0 |     n = nmap.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0 |     o = warezmaster.
   32     0     0     0     0     0     0     0     0     0     0     0     0     0     0    70     0 |     p = warezclient.
    1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     q = rootkit.
```

1) **normal:** la TP Rate risulta essere la peggiore dopo quella del Naive Bayes
2) **buffer_overflow:** come gli altri algoritmi (nulla)
3) **neptune, smurf, pod:** classi in cui è stata ottenuta una precisione del 100%
4) **guess_password:** risultato migliore di J48, ma peggiore degli altri
5) **back:** unico algoritmo ad aver ottenuto risultati negativi nella classificazione di queste istanze
6) **satan:** migliore solo del Naive Bayes
7) **nmap:** prestazioni peggiori
8) **warezclient:** prestazioni peggiori
9) **imap, rootkit:** come tutti gli altri algoritmi (0%)

Considerazioni personali: l*a rete neurale non ha prodotto risultati particolarmente interessanti. In generale la classificherei migliore solo del Naive Bayes, ovviamente non tenendo conto dell' enormità di tempo richiesto per il suo addestramento*