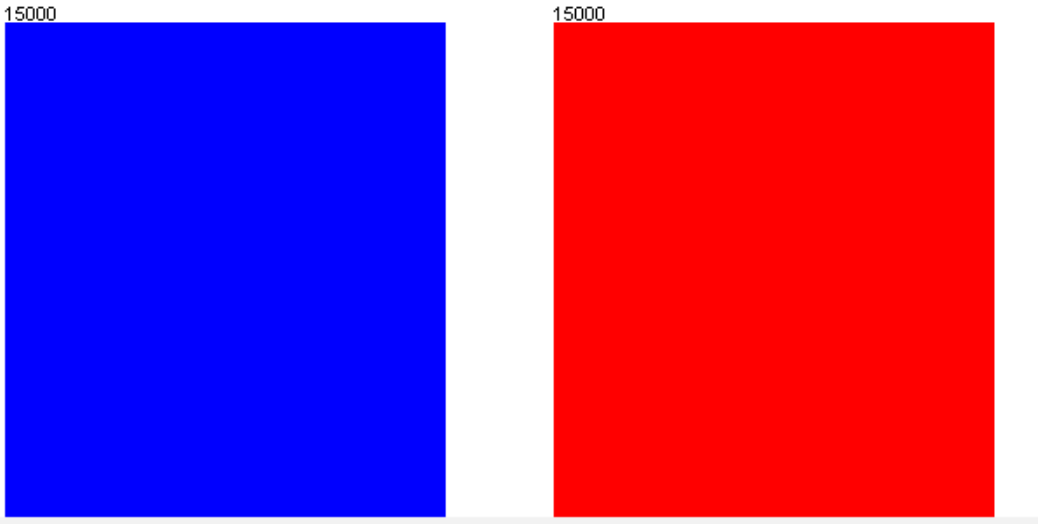
BINARNA KLASIFIKACIJA

Selected attribute

Name: classification Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	BENIGN	15000	15000
2	ATTACK	15000	15000

Class: classification (Nom) Visualize All



Current relation

Relation: sampled_train_bin Attributes: 79
Instances: 30000 Sum of weights: 30000

Naive Bayes

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

Correctly Classified Instances       24170               80.5667 %
Incorrectly Classified Instances      5830               19.4333 %
Kappa statistic                          0.6113
Mean absolute error                      0.1978
Root mean squared error                  0.433
Relative absolute error                 39.5661 %
Root relative squared error             86.5979 %
Total Number of Instances            30000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.644    0.033    0.951      0.644    0.768      0.646    0.888     0.888     BENIGN
                 0.967    0.356    0.731      0.967    0.833      0.646    0.879     0.841     ATTACK
Weighted Avg.    0.806    0.194    0.841      0.806    0.800      0.646    0.884     0.864

=== Confusion Matrix ===

     a     b   <-- classified as
  9665  5335 |     a = BENIGN
   495 14505 |     b = ATTACK
```

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     sampled_test_bin
Instances:     unknown (yet). Reading incrementally
Attributes:   79

=== Summary ===

Correctly Classified Instances      179862               80.7922 %
Incorrectly Classified Instances     42761               19.2078 %
Kappa statistic                          0.6158
Mean absolute error                      0.1978
Root mean squared error                  0.4329
Total Number of Instances           222623

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.649    0.034    0.951      0.649    0.772      0.649    0.887     0.885     BENIGN
                 0.966    0.351    0.734      0.966    0.834      0.649    0.878     0.840     ATTACK
Weighted Avg.    0.808    0.192    0.842      0.808    0.803      0.649    0.882     0.862

=== Confusion Matrix ===

      a      b   <-- classified as
  72291  39021 |      a = BENIGN
   3740 107571 |      b = ATTACK
```

Kod ovog algoritma imamo oko 80% tačno klasifikovanih instanci. Preciznost klasifikacije napada (0,731 i 0,734) je manja u odnosu na preciznost klasifikacije čistih instanci (0,951). Do ovoga je došlo jer je sa većom verovatnoćom lakše uočiti patern benignih instanci.

Normalized na vrednosti od –1 do 1

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

Correctly Classified Instances       23776               79.2533 %
Incorrectly Classified Instances      6224               20.7467 %
Kappa statistic                          0.5851
Mean absolute error                      0.2075
Root mean squared error                  0.443
Relative absolute error                 41.4941 %
Root relative squared error             88.5972 %
Total Number of Instances            30000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.622     0.037     0.944       0.622   0.750      0.622    0.873     0.875     BENIGN
               0.963     0.378     0.718       0.963   0.823      0.622    0.864     0.829     ATTACK
Weighted Avg.  0.793     0.207     0.831       0.793   0.786      0.622    0.868     0.852

=== Confusion Matrix ===

     a     b   <-- classified as
  9328  5672 |     a = BENIGN
   552 14448 |     b = ATTACK
```

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     sampled_test_bin
Instances:     unknown (yet). Reading incrementally
Attributes:   79

=== Summary ===

Correctly Classified Instances      136617               61.367  %
Incorrectly Classified Instances     86006               38.633  %
Kappa statistic                          0.2273
Mean absolute error                      0.3859
Root mean squared error                  0.6204
Total Number of Instances           222623

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.785     0.558     0.585       0.785   0.670      0.242    0.638     0.581     BENIGN
               0.442     0.215     0.673       0.442   0.534      0.242    0.502     0.579     ATTACK
Weighted Avg.  0.614     0.386     0.629       0.614   0.602      0.242    0.570     0.580

=== Confusion Matrix ===

     a     b   <-- classified as
 87379 23933 |     a = BENIGN
 62073 49238 |     b = ATTACK
```

Pokušala sam da odradim normalizaciju vrednosti atributa na opseg od –1 do 1, što je samo pogoršalo rezultate jer su podaci već bili proporcionalno normalizovani u preprocessing-u. Smanjenje opsega nije donelo ništa dobro. Tačnost klasifikacije spala je na 61.367%, odnosno srednja kvadratna greška je 0,6204.

Random Forest

Classifier output

```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 7.47 seconds

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

Correctly Classified Instances       29915               99.7167 %
Incorrectly Classified Instances        85                0.2833 %
Kappa statistic                          0.9943
Mean absolute error                      0.0073
Root mean squared error                  0.0483
Relative absolute error                  1.4634 %
Root relative squared error              9.6594 %
Total Number of Instances            30000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.004    0.996      0.998    0.997      0.994    1.000     1.000     BENIGN
                 0.996    0.002    0.998      0.996    0.997      0.994    1.000     1.000     ATTACK
Weighted Avg.    0.997    0.003    0.997      0.997    0.997      0.994    1.000     1.000

=== Confusion Matrix ===

     a     b   <-- classified as
 14972    28 |     a = BENIGN
    57 14943 |     b = ATTACK
```

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     sampled_test_bin
Instances:     unknown (yet). Reading incrementally
Attributes:   79

=== Summary ===

Correctly Classified Instances      222046               99.7408 %
Incorrectly Classified Instances       577                0.2592 %
Kappa statistic                          0.9948
Mean absolute error                      0.0072
Root mean squared error                  0.0477
Total Number of Instances           222623

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.003    0.997      0.998    0.997      0.995    1.000     1.000     BENIGN
                 0.997    0.002    0.998      0.997    0.997      0.995    1.000     1.000     ATTACK
Weighted Avg.    0.997    0.003    0.997      0.997    0.997      0.995    1.000     1.000

=== Confusion Matrix ===

      a      b   <-- classified as
 111124    188 |      a = BENIGN
    389 110922 |      b = ATTACK
```

Kod ovog algoritma rezultati su značajno bolji. Tačnost klasifikacije je na 99.7%, što je veoma zadovoljavajuće za ovu vrstu podataka. U ovom slučaju bolje su klasifikovani napadi u odnosu na čiste podatke, ali razlike su minimalne. Preciznost je 0,997,a srednja kvadratna greška pala je na 0,0477. Više je benignih podataka koji su klasifikovani kao napadi netačno nego što je napada koji su klasifikovani kao benigni podaci.

Tree J48

Classifier output

```
Number of Leaves  :     74

Size of the tree :      147


Time taken to build model: 2.24 seconds

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

Correctly Classified Instances       29907               99.69   %
Incorrectly Classified Instances        93                0.31   %
Kappa statistic                          0.9938
Mean absolute error                      0.0044
Root mean squared error                  0.0543
Relative absolute error                  0.8716 %
Root relative squared error             10.8541 %
Total Number of Instances            30000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.003    0.997      0.996    0.997      0.994    0.998     0.998     BENIGN
                 0.997    0.004    0.996      0.997    0.997      0.994    0.998     0.997     ATTACK
Weighted Avg.    0.997    0.003    0.997      0.997    0.997      0.994    0.998     0.997

=== Confusion Matrix ===

     a     b   <-- classified as
 14947    53 |     a = BENIGN
    40 14960 |     b = ATTACK
```

Ovaj algoritam je malo slabije klasifikovao instance (99.69% i 99,65% tačnih klasifikacija) u odnosu na Random Forest, ali značajno bolje od Naive Bayes algoritma. Preciznost je neznatno veća na benignim podacima nego na napadima za razliku od Random Forest algoritma.

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     sampled_test_bin
Instances:     unknown (yet). Reading incrementally
Attributes:   79

=== Summary ===

Correctly Classified Instances      221835               99.646  %
Incorrectly Classified Instances       788                0.354  %
Kappa statistic                          0.9929
Mean absolute error                      0.0046
Root mean squared error                  0.0577
Total Number of Instances           222623

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.002    0.998      0.995    0.996      0.993    0.998     0.998     BENIGN
                 0.998    0.005    0.995      0.998    0.996      0.993    0.998     0.996     ATTACK
Weighted Avg.    0.996    0.004    0.996      0.996    0.996      0.993    0.998     0.997

=== Confusion Matrix ===

      a      b   <-- classified as
 110772    540 |      a = BENIGN
    248 111063 |      b = ATTACK
```

Unpruned =TRUE

Ovde, za razliku od višeklasne klasifikacije imamo bolje rezultate kada je isključen pruning (99.71% i 99.6698%). Stablo odlučivanja koje nije orezano može preciznije reproducirati trening podatke, posebno ako postoji puno specifičnosti ili kompleksnosti u podacima. To može dovesti do boljih rezultata na trening skupu, jer se model može potpuno prilagoditi karakteristikama trening podataka. Preciznost je ostala idalje malo veća kod klasifikacije benignih podataka.

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     sampled_test_bin
Instances:     unknown (yet). Reading incrementally
Attributes:   79

=== Summary ===

Correctly Classified Instances      221888               99.6698 %
Incorrectly Classified Instances       735                0.3302 %
Kappa statistic                          0.9934
Mean absolute error                      0.0041
Root mean squared error                  0.0559
Total Number of Instances           222623

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.002    0.998      0.995    0.997      0.993    0.997     0.997     BENIGN
                 0.998    0.005    0.996      0.998    0.997      0.993    0.997     0.994     ATTACK
Weighted Avg.    0.997    0.003    0.997      0.997    0.997      0.993    0.997     0.996

=== Confusion Matrix ===

      a      b   <-- classified as
 110811    501 |      a = BENIGN
    234 111077 |      b = ATTACK
```

Classifier output

```
Number of Leaves  :     88

Size of the tree :      175


Time taken to build model: 1.8 seconds

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

Correctly Classified Instances       29913               99.71   %
Incorrectly Classified Instances        87                0.29   %
Kappa statistic                          0.9942
Mean absolute error                      0.0037
Root mean squared error                  0.0515
Relative absolute error                  0.7369 %
Root relative squared error             10.3077 %
Total Number of Instances            30000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.002    0.998      0.996    0.997      0.994    0.998     0.998     BENIGN
                 0.998    0.004    0.996      0.998    0.997      0.994    0.998     0.996     ATTACK
Weighted Avg.    0.997    0.003    0.997      0.997    0.997      0.994    0.998     0.997

=== Confusion Matrix ===

     a     b   <-- classified as
 14943    57 |     a = BENIGN
    30 14970 |     b = ATTACK
```

OneR algoritam

OneR je jednostavan klasifikator koji bira jedan atribut i za njega definiše pravila na osnovu kojih se radi klasifikacija.

```
=== Classifier model (full training set) ===

_Destination_Port:
        < -0.3862220806224067   -> BENIGN
        < -0.38296680237107955  -> ATTACK
        < -0.38228996233862533  -> BENIGN
        < -0.3806784384518297   -> ATTACK
```

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

Correctly Classified Instances       27331               91.1033 %
Incorrectly Classified Instances      2669                8.8967 %
Kappa statistic                          0.8221
Mean absolute error                      0.089
Root mean squared error                  0.2983
Relative absolute error                 17.7933 %
Root relative squared error             59.6546 %
Total Number of Instances            30000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.840    0.018    0.979      0.840    0.904      0.831    0.911     0.903     BENIGN
                 0.982    0.160    0.860      0.982    0.917      0.831    0.911     0.853     ATTACK
Weighted Avg.    0.911    0.089    0.920      0.911    0.911      0.831    0.911     0.878

=== Confusion Matrix ===

     a     b   <-- classified as
 12598  2402 |     a = BENIGN
   267 14733 |     b = ATTACK
```

I ovaj algoritam je dao značajno bolje rezultate u odnosu na Naive Bayes algoritam, ali nešto slabije rezultate u odnosu na Tree algoritme. Tačno klasifikovanih instanci je 91.1033% (cross-validation) odnosno 91.2004% (test set). Sa većom preciznošću su i ovde klasifikovane benigne instance (0,979 i 0,980 naspram 0,860 i 0,861).

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     sampled_test_bin
Instances:     unknown (yet). Reading incrementally
Attributes:   79

=== Summary ===

Correctly Classified Instances      203033               91.2004 %
Incorrectly Classified Instances     19590                8.7996 %
Kappa statistic                          0.824
Mean absolute error                      0.088
Root mean squared error                  0.2966
Total Number of Instances           222623

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.841    0.017    0.980      0.841    0.905      0.832    0.912     0.904     BENIGN
                 0.983    0.159    0.861      0.983    0.918      0.832    0.912     0.855     ATTACK
Weighted Avg.    0.912    0.088    0.920      0.912    0.912      0.832    0.912     0.879

=== Confusion Matrix ===

     a      b   <-- classified as
 93643  17669 |      a = BENIGN
  1921 109390 |      b = ATTACK
```

K Nearest Neighbors (kNN) -IBk

Cross-validation KNN=1

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

Correctly Classified Instances       29563               98.5433 %
Incorrectly Classified Instances       437                1.4567 %
Kappa statistic                          0.9709
Mean absolute error                      0.0146
Root mean squared error                  0.1207
Relative absolute error                  2.9202 %
Root relative squared error             24.1376 %
Total Number of Instances            30000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.013    0.987      0.983    0.985      0.971    0.986     0.980     BENIGN
                 0.987    0.017    0.984      0.987    0.985      0.971    0.986     0.979     ATTACK
Weighted Avg.    0.985    0.015    0.985      0.985    0.985      0.971    0.986     0.979

=== Confusion Matrix ===

     a     b   <-- classified as
 14752   248 |     a = BENIGN
   189 14811 |     b = ATTACK
```

Test set KNN=1

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     sampled_test_bin
Instances:     unknown (yet). Reading incrementally
Attributes:   79

=== Summary ===

Correctly Classified Instances      219272               98.4948 %
Incorrectly Classified Instances      3351                1.5052 %
Kappa statistic                          0.9699
Mean absolute error                      0.0151
Root mean squared error                  0.1227
Total Number of Instances           222623

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.013    0.987      0.983    0.985      0.970    0.986     0.979     BENIGN
                 0.987    0.017    0.983      0.987    0.985      0.970    0.986     0.978     ATTACK
Weighted Avg.    0.985    0.015    0.985      0.985    0.985      0.970    0.986     0.979

=== Confusion Matrix ===

      a      b   <-- classified as
 109416   1896 |      a = BENIGN
   1455 109856 |      b = ATTACK
```

Cross-validation KNN=7

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

Correctly Classified Instances       29570               98.5667 %
Incorrectly Classified Instances       430                1.4333 %
Kappa statistic                          0.9713
Mean absolute error                      0.0199
Root mean squared error                  0.1079
Relative absolute error                  3.9776 %
Root relative squared error             21.5739 %
Total Number of Instances            30000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.979    0.007    0.993      0.979    0.986      0.971    0.996     0.995     BENIGN
                 0.993    0.021    0.979      0.993    0.986      0.971    0.996     0.993     ATTACK
Weighted Avg.    0.986    0.014    0.986      0.986    0.986      0.971    0.996     0.994

=== Confusion Matrix ===

     a     b   <-- classified as
 14678   322 |     a = BENIGN
   108 14892 |     b = ATTACK
```

Test set KNN=7

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     sampled_test_bin
Instances:     unknown (yet). Reading incrementally
Attributes:   79

=== Summary ===

Correctly Classified Instances      219031               98.3865 %
Incorrectly Classified Instances      3592                1.6135 %
Kappa statistic                          0.9677
Mean absolute error                      0.0216
Root mean squared error                  0.1139
Total Number of Instances           222623

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.977    0.010    0.990      0.977    0.984      0.968    0.995     0.995     BENIGN
                 0.990    0.023    0.978      0.990    0.984      0.968    0.995     0.992     ATTACK
Weighted Avg.    0.984    0.016    0.984      0.984    0.984      0.968    0.995     0.993

=== Confusion Matrix ===

      a      b   <-- classified as
 108789   2523 |      a = BENIGN
   1069 110242 |      b = ATTACK
```

Ovaj algoritam je klasifikovao instance bolje od Naive Bayes algoritma i OneR alogoritma, ali nešto slabije od Tree algoritama. Tačno klasifikovanih ima 98,5433% kada je k=1, odnosno 98,5667% kada je k=7 pri cross-validaciji. Kada su u pitanju test setovi podataka ovi procenti iznose 98.4948% i 98.3865% respektivno. Uočava se da kada je u pitanju povećanje broja komšija kod cross-validation imamo veći procenat tačno klasifikovanih, dok je kod test seta ovo suprotno. I kod ovog algoritma je preciynost klasifikovanja benignih instanci nešto veća nego kad su napadi u pitanju.