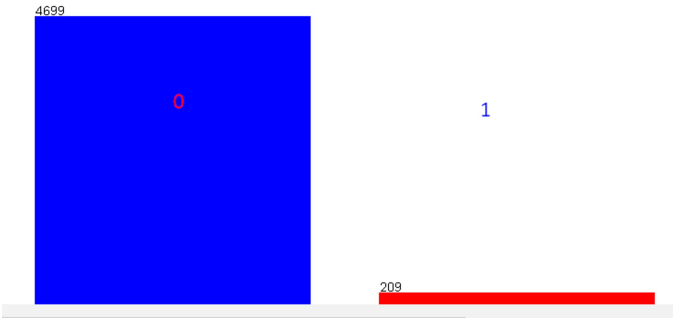
Skladištenje

1. Sklonili smo 'Other' vrednost iz atributa 'Gender' u dataset-u.Postojala je samo 1 instanca sa tom vrednošću.
2. Sve vrednosti koje su bile N/A za atribut 'bmi' smo stavili da budu 0 jer weka nije htela da ih prikaže, nije htelo da izbaci.Onda smo ih izbacili. Promenjen je tip atributa sa string na numeric preko filtera StringToNumeric.
3. Stavljeno je da hypertension, heart_disease i stroke budu ili 0 ili 1, umesto opsega [0,1] preko filtera NumericToNominal.
4. Atribut 'ID' smo uklonili zato što je bio nepotreban.

Odnos 0 i 1 u Stroke klasi(0 - individue koje nisu imale šlog, 1-oni koji jesu):



Na samom početku smo probali neke algoritme, čisto radi eksperimentisanja, i videli da za skoro polovinu algoritama je klasifikacija bila loša iz razloga što sve podatke stavi samo u jednu klasu (klasu a), i na osnovu toga utvrdili da je potrebno izbalansirati dataset.

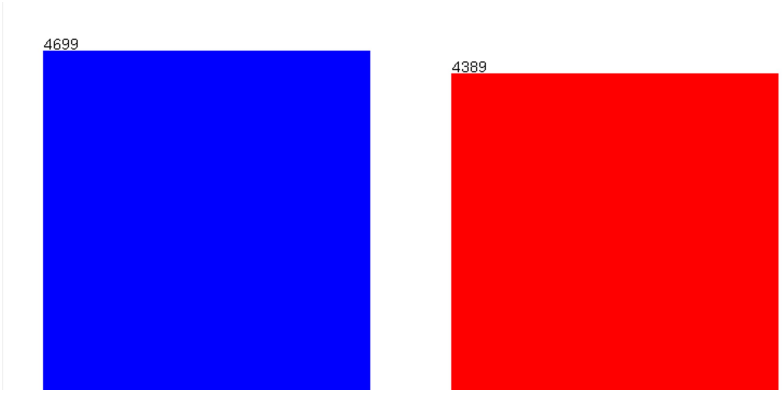
Na kraju smo se vratili da pustimo Random Forest nad originalnim nebalansiranim dataset-om radi poređenja rezultata.

```
Time taken to build model: 0.05 seconds

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

Correctly Classified Instances        4381               89.2624 %
Incorrectly Classified Instances       527               10.7376 %
Kappa statistic                          0.1681
Mean absolute error                      0.1235
Root mean squared error                  0.2808
Relative absolute error                151.1293 %
Root relative squared error            139.0914 %
Total Number of Instances             4908

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.917    0.651    0.969      0.917    0.942      0.184    0.826     0.991     0
                 0.349    0.083    0.157      0.349    0.217      0.184    0.826     0.167     1
Weighted Avg.    0.893    0.627    0.935      0.893    0.911      0.184    0.826     0.955

=== Confusion Matrix ===

    a    b   <-- classified as
 4308  391 |    a = 0
  136   73 |    b = 1
```

Tehnike balansiranja

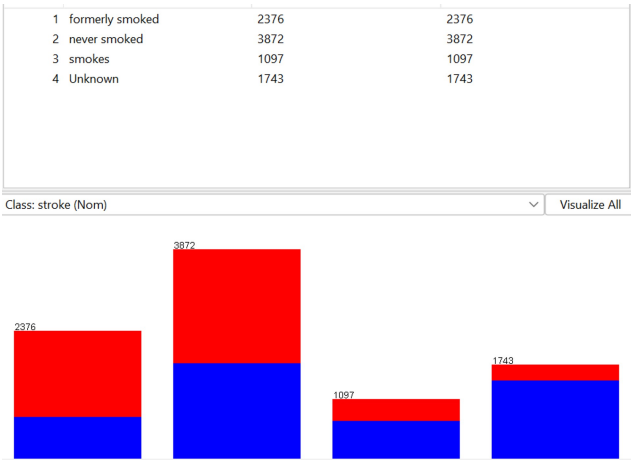
Balansiranje podataka smo radili na dva načina (Oversampling, Undersampling).

Oversample-ovanje se radilo da se izbalansiraju podaci, da se dovede na broj ljudi kod kojih je odnos da/ne za moždani udar otprilike 50-50. Na slici ispod je Oversampling odrađen odokativno.



Oversampling je odrađen pomoću filtera SMOTE. Procenat za koji se povećava broj instanci klase b je stavljen na 2000.

Mala digresija, zanimljivi su ovi podaci za razmatranje, najgori su ovi koji su formerly smoked, a iznenađujući su podaci koji se odnose na never smoked (dosta njih ima moždani udar) i smokes (poprilično mali broj je pozitivan u pogledu moždanog udara):



Radi se cross-validation za broj foldova 10:

- Naive Bayes pušten za te podatke. On obezbeđuje maksimizaciju verovatnoće ako su karakteristike veoma nezavisne, tj ako su atributi međusobno nezavisni (međutim, iako je ovaj algoritam brz, ne možemo nikako gledati atribute za ovaj slučaj kao nezavisne):

```
=== Summary ===

Correctly Classified Instances        7375               81.151  %
Incorrectly Classified Instances      1713               18.849  %
Kappa statistic                          0.6248
Mean absolute error                      0.217
Root mean squared error                  0.3694
Relative absolute error                 43.4438 %
Root relative squared error             73.9214 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
```

```
=== Summary ===

Correctly Classified Instances        7375               81.151  %
Incorrectly Classified Instances      1713               18.849  %
Kappa statistic                          0.6248
Mean absolute error                      0.217
Root mean squared error                  0.3694
Relative absolute error                 43.4438 %
Root relative squared error             73.9214 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.734    0.106    0.881      0.734    0.801      0.634    0.896     0.919     0
                 0.894    0.266    0.759      0.894    0.821      0.634    0.896     0.859     1
Weighted Avg.    0.812    0.183    0.822      0.812    0.811      0.634    0.896     0.890

=== Confusion Matrix ===

    a    b   <-- classified as
 3451 1248 |    a = 0
  465 3924 |    b = 1
```

- Random Forest
  Objedinjuje mnoga nezavisna stabla odlučivanja i, ponovnim uzorkovanjem, kreira različite podskupove instanci za obavljanje klasifikacije. Svako stablo odlučivanja daje sopstveni rezultat klasifikacije, a zatim se konačna klasa izvodi putem većinskog glasanja.

  Daje dosta fine podatke u smislu F-score-a, prelazi 0.95 što je odlično, a ROC kriva mu je takođe dobra.

```
=== Summary ===

Correctly Classified Instances        8662               95.3125 %
Incorrectly Classified Instances       426                4.6875 %
Kappa statistic                          0.9062
Mean absolute error                      0.0898
Root mean squared error                  0.1967
Relative absolute error                 17.9717 %
Root relative squared error             39.3682 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.951    0.044    0.958      0.951    0.954      0.906    0.987     0.986     0
                 0.956    0.049    0.948      0.956    0.952      0.906    0.987     0.988     1
Weighted Avg.    0.953    0.047    0.953      0.953    0.953      0.906    0.987     0.987

=== Confusion Matrix ===

    a    b   <-- classified as
 4468  231 |    a = 0
  195 4194 |    b = 1
```

- Algoritam KNN smo isprobali za različite vrednosti K, i dobili sledeće rezultate:

  K=1

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

Correctly Classified Instances        8484               93.3539 %
```

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

Correctly Classified Instances        8484               93.3539 %
Incorrectly Classified Instances       604                6.6461 %
Kappa statistic                          0.8671
Mean absolute error                      0.0666
Root mean squared error                  0.2578
Relative absolute error                 13.329  %
Root relative squared error             51.5839 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.914    0.045    0.956      0.914    0.934      0.868    0.934     0.921     0
                 0.955    0.086    0.912      0.955    0.933      0.868    0.934     0.894     1
Weighted Avg.    0.934    0.065    0.935      0.934    0.934      0.868    0.934     0.908

=== Confusion Matrix ===

    a    b   <-- classified as
 4293  406 |    a = 0
  198 4191 |    b = 1
```

K=3

```
=== Summary ===

Correctly Classified Instances        8539               93.9591 %
Incorrectly Classified Instances       549                6.0409 %
Kappa statistic                          0.8792
Mean absolute error                      0.0763
Root mean squared error                  0.2346
Relative absolute error                 15.2829 %
Root relative squared error             46.9482 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.919    0.038    0.963      0.919    0.940      0.880    0.959     0.949     0
                 0.962    0.081    0.917      0.962    0.939      0.880    0.959     0.935     1
Weighted Avg.    0.940    0.059    0.941      0.940    0.940      0.880    0.959     0.943

=== Confusion Matrix ===

    a    b   <-- classified as
 4317  382 |    a = 0
  167 4222 |    b = 1
```

K=5

```
=== Summary ===

Correctly Classified Instances        8497               93.4969 %
Incorrectly Classified Instances       591                6.5031 %
Kappa statistic                          0.87
Mean absolute error                      0.0824
Root mean squared error                  0.235
Relative absolute error                 16.5066 %
Root relative squared error             47.0228 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.909    0.037    0.963      0.909    0.935      0.872    0.964     0.955     0
                 0.963    0.091    0.908      0.963    0.935      0.872    0.964     0.944     1
Weighted Avg.    0.935    0.063    0.937      0.935    0.935      0.872    0.964     0.950

=== Confusion Matrix ===

    a    b   <-- classified as
 4270  429 |    a = 0
  162 4227 |    b = 1
```

K=7

```
=== Summary ===

Correctly Classified Instances        8462               93.1118 %
Incorrectly Classified Instances       626                6.8882 %
Kappa statistic                          0.8624
Mean absolute error                      0.0874
Root mean squared error                  0.2384
Relative absolute error                 17.4905 %
Root relative squared error             47.7176 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.900    0.036    0.964      0.900    0.931      0.864    0.966     0.959     0
                 0.964    0.100    0.900      0.964    0.931      0.864    0.966     0.949     1
Weighted Avg.    0.931    0.067    0.933      0.931    0.931      0.864    0.966     0.954

=== Confusion Matrix ===

    a    b   <-- classified as
 4229  470 |    a = 0
  156 4233 |    b = 1
```

- Logistic Regression algoritam procenjuje verovatnoću da instanca pripada pozitivnoj klasi (označena kao klasa a) koristeći logističku funkciju, takođe poznatu kao sigmoidna funkcija.

```
=== Summary ===

Correctly Classified Instances        7547               83.0436 %
Incorrectly Classified Instances      1541               16.9564 %
Kappa statistic                          0.661
Mean absolute error                      0.2296
Root mean squared error                  0.3385
Relative absolute error                 45.982  %
Root relative squared error             67.7478 %
```

```
=== Summary ===

Correctly Classified Instances        7547               83.0436 %
Incorrectly Classified Instances      1541               16.9564 %
Kappa statistic                          0.661
Mean absolute error                      0.2296
Root mean squared error                  0.3385
Relative absolute error                 45.982  %
Root relative squared error             67.7478 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.815    0.153    0.851      0.815    0.833      0.662    0.917     0.932     0
                 0.847    0.185    0.811      0.847    0.828      0.662    0.917     0.894     1
Weighted Avg.    0.830    0.168    0.831      0.830    0.830      0.662    0.917     0.913

=== Confusion Matrix ===

    a    b   <-- classified as
 3830  869 |    a = 0
  672 3717 |    b = 1
```

- J48 Decision tree

Podešavanja J48 su bila sledeća: faktor pouzdanosti je postavljen na 0.25, a unpruned je postavljen na false
Vizualizacija stabla je nemoguća, odnosno mora da se mnogo zumira.

```
=== Summary ===

Correctly Classified Instances        8458               93.0678 %
Incorrectly Classified Instances       630                6.9322 %
Kappa statistic                          0.8613
Mean absolute error                      0.0965
Root mean squared error                  0.2526
Relative absolute error                 19.3183 %
Root relative squared error             50.5434 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.927    0.066    0.938      0.927    0.933      0.861    0.943     0.934     0
                 0.934    0.073    0.923      0.934    0.929      0.861    0.943     0.907     1
Weighted Avg.    0.931    0.069    0.931      0.931    0.931      0.861    0.943     0.921

=== Confusion Matrix ===

    a    b   <-- classified as
 4357  342 |    a = 0
  288 4101 |    b = 1
```

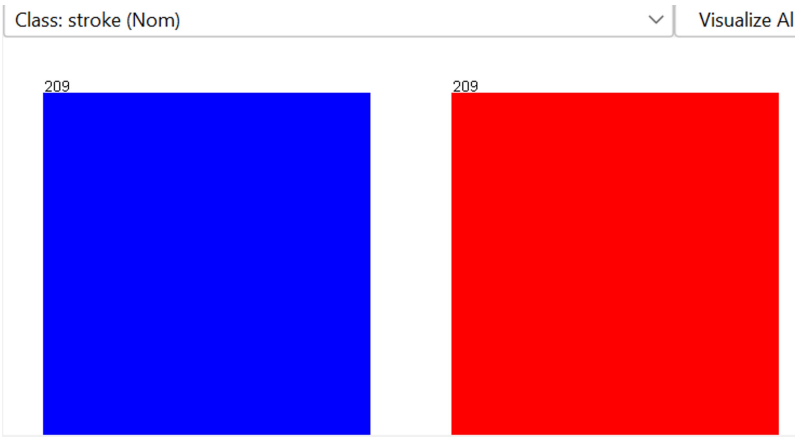
- SGD

U svakoj iteraciji izračunava gradijent koristeći jedan uzorak. Omogućava minibatch i stoga je pogodan za velike probleme.

```
=== Summary ===

Correctly Classified Instances        7512               82.6585 %
Incorrectly Classified Instances      1576               17.3415 %
Kappa statistic                          0.653
Mean absolute error                      0.1734
Root mean squared error                  0.4164
Relative absolute error                 34.7235 %
Root relative squared error             83.3349 %
Total Number of Instances             9088

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.821    0.167    0.840      0.821    0.830      0.653    0.827     0.782     0
                 0.833    0.179    0.813      0.833    0.823      0.653    0.827     0.758     1
Weighted Avg.    0.827    0.173    0.827      0.827    0.827      0.653    0.827     0.770

=== Confusion Matrix ===

    a    b   <-- classified as
 3858  841 |    a = 0
  735 3654 |    b = 1
```

Undersample-ovanje se radilo da se izbalansiraju podaci na random undersampling, odnosno da se dovede na broj ljudi kod kojih je odnos da/ne za moždani udar tačno 50-50.

Ovo se uradilo tako što se za filter odabere SpreadSubsample, distributionSpread -> 1.0



Probali smo iste algoritme i za Undersampling i uočili da daje dosta gore rezultate od Oversampling-a.

- Naive Bayes

```
=== Summary ===

Correctly Classified Instances         302               72.2488 %
Incorrectly Classified Instances       116               27.7512 %
Kappa statistic                          0.445
Mean absolute error                      0.2971
Root mean squared error                  0.4388
Relative absolute error                 59.4243 %
Root relative squared error             87.7545 %
Total Number of Instances              418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.651    0.206    0.760      0.651    0.701      0.450    0.809     0.836     0
                 0.794    0.349    0.695      0.794    0.741      0.450    0.809     0.781     1
Weighted Avg.    0.722    0.278    0.727      0.722    0.721      0.450    0.809     0.809

=== Confusion Matrix ===

   a   b   <-- classified as
 136  73 |   a = 0
  43 166 |   b = 1
```

- SGD

```
=== Summary ===

Correctly Classified Instances         302               72.2488 %
Incorrectly Classified Instances       116               27.7512 %
Kappa statistic                          0.445
Mean absolute error                      0.2775
Root mean squared error                  0.5268
Relative absolute error                 55.5017 %
Root relative squared error            105.3574 %
Total Number of Instances              418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.679    0.234    0.743      0.679    0.710      0.447    0.722     0.665     0
                 0.766    0.321    0.705      0.766    0.734      0.447    0.722     0.657     1
Weighted Avg.    0.722    0.278    0.724      0.722    0.722      0.447    0.722     0.661

=== Confusion Matrix ===

   a   b   <-- classified as
 142  67 |   a = 0
  49 160 |   b = 1
```

- KNN

K=1

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

Correctly Classified Instances         271               64.8325 %
Incorrectly Classified Instances       147               35.1675 %
Kappa statistic                          0.2967
Mean absolute error                      0.3525
Root mean squared error                  0.5915
Relative absolute error                 70.4909 %
Root relative squared error            118.2899 %
Total Number of Instances              418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.670    0.373    0.642      0.670    0.656      0.297    0.655     0.606     0
                 0.627    0.330    0.655      0.627    0.641      0.297    0.655     0.607     1
Weighted Avg.    0.648    0.352    0.649      0.648    0.648      0.297    0.655     0.607

=== Confusion Matrix ===

   a   b   <-- classified as
 140  69 |   a = 0
  78 131 |   b = 1
```

K=3

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

Correctly Classified Instances         265               63.3971 %
Incorrectly Classified Instances       153               36.6029 %
Kappa statistic                          0.2679
Mean absolute error                      0.3941
Root mean squared error                  0.5029
Relative absolute error                 78.8244 %
Root relative squared error            100.5688 %
Total Number of Instances              418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.670    0.402    0.625      0.670    0.647      0.269    0.673     0.668     0
                 0.598    0.330    0.644      0.598    0.620      0.269    0.673     0.623     1
Weighted Avg.    0.634    0.366    0.635      0.634    0.633      0.269    0.673     0.646

=== Confusion Matrix ===

   a   b   <-- classified as
 140  69 |   a = 0
  84 125 |   b = 1
```

K=5

```
=== Summary ===

Correctly Classified Instances         272               65.0718 %
Incorrectly Classified Instances       146               34.9282 %
Kappa statistic                          0.3014
Mean absolute error                      0.3963
Root mean squared error                  0.4782
Relative absolute error                 79.2555 %
Root relative squared error             95.632  %
Total Number of Instances              418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.636    0.335    0.655      0.636    0.646      0.302    0.697     0.720     0
                 0.665    0.364    0.647      0.665    0.656      0.302    0.697     0.635     1
Weighted Avg.    0.651    0.349    0.651      0.651    0.651      0.302    0.697     0.677

=== Confusion Matrix ===

   a   b   <-- classified as
 133  76 |   a = 0
  70 139 |   b = 1
```

K=7

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

Correctly Classified Instances         278               66.5072 %
Incorrectly Classified Instances       140               33.4928 %
Kappa statistic                          0.3301
Mean absolute error                      0.3996
Root mean squared error                  0.4679
Relative absolute error                 79.9185 %
Root relative squared error             93.5816 %
Total Number of Instances              418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.651    0.321    0.670      0.651    0.660      0.330    0.706     0.724     0
                 0.679    0.349    0.660      0.679    0.670      0.330    0.706     0.661     1
Weighted Avg.    0.665    0.335    0.665      0.665    0.665      0.330    0.706     0.693

=== Confusion Matrix ===

   a   b   <-- classified as
 136  73 |   a = 0
  67 142 |   b = 1
```

Ono što je interesantno je da se za veće K povećava ROC kriva, a negde je F-score bolji.

- Random Forest

```
=== Summary ===

Correctly Classified Instances         304               72.7273 %
Incorrectly Classified Instances       114               27.2727 %
Kappa statistic                          0.4545
Mean absolute error                      0.3451
Root mean squared error                  0.4284
Relative absolute error                 69.0225 %
Root relative squared error             85.6725 %
Total Number of Instances               418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.689    0.234    0.746      0.689    0.716      0.456    0.795     0.821     0
                 0.766    0.311    0.711      0.766    0.737      0.456    0.795     0.754     1
Weighted Avg.    0.727    0.273    0.729      0.727    0.727      0.456    0.795     0.788

=== Confusion Matrix ===

   a   b   <-- classified as
 144  65 |   a = 0
  49 160 |   b = 1
```

- J48

```
=== Summary ===

Correctly Classified Instances         306               73.2057 %
Incorrectly Classified Instances       112               26.7943 %
Kappa statistic                          0.4641
Mean absolute error                      0.3245
Root mean squared error                  0.4587
Relative absolute error                 64.8985 %
Root relative squared error             91.7307 %
Total Number of Instances               418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.689    0.225    0.754      0.689    0.720      0.466    0.748     0.729     0
                 0.775    0.311    0.714      0.775    0.743      0.466    0.748     0.692     1
Weighted Avg.    0.732    0.268    0.734      0.732    0.732      0.466    0.748     0.710

=== Confusion Matrix ===

   a   b   <-- classified as
 144  65 |   a = 0
  47 162 |   b = 1
```

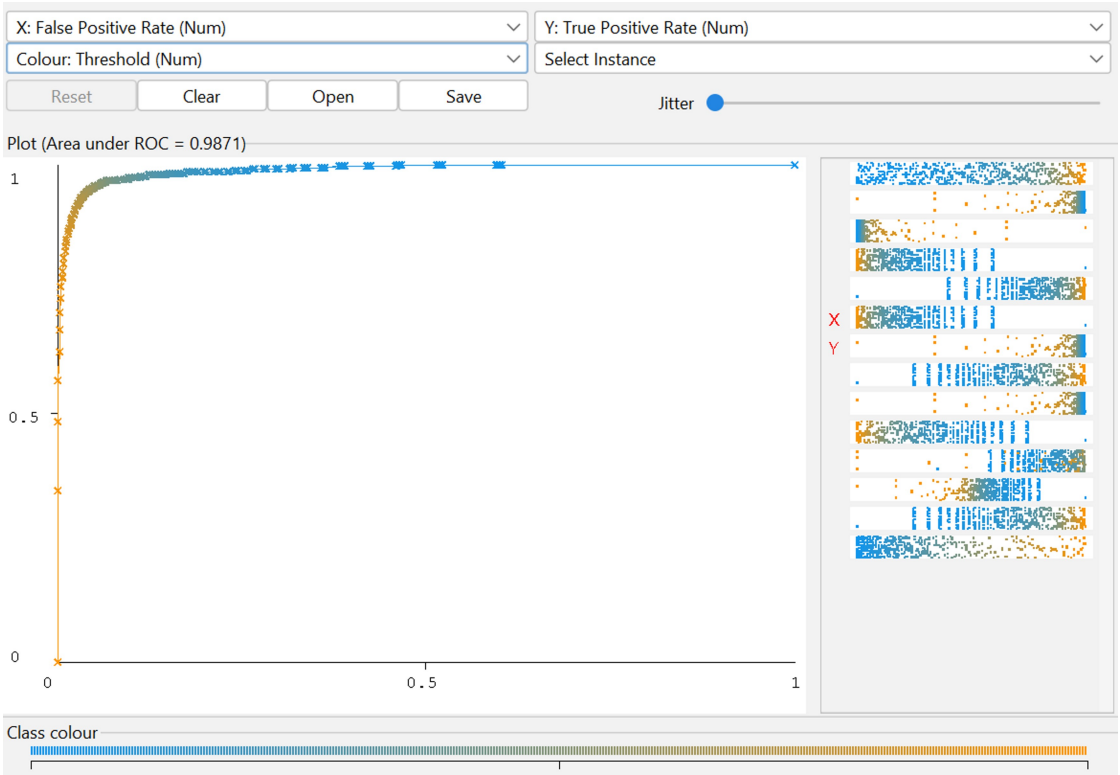
- Logistic Regression

```
=== Summary ===

Correctly Classified Instances         304               72.7273 %
Incorrectly Classified Instances       114               27.2727 %
Kappa statistic                          0.4545
Mean absolute error                      0.3354
Root mean squared error                  0.4204
Relative absolute error                 67.0713 %
Root relative squared error             84.0822 %
Total Number of Instances              418

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.689    0.234    0.746      0.689    0.716      0.456    0.810     0.826     0
                 0.766    0.311    0.711      0.766    0.737      0.456    0.810     0.771     1
Weighted Avg.    0.727    0.273    0.729      0.727    0.727      0.456    0.810     0.798

=== Confusion Matrix ===

   a   b   <-- classified as
 144  65 |   a = 0
  49 160 |   b = 1
```

Kada smo došli do znanja da je Oversampling bolje rešenje za predviđanje i da Random Forest daje najbolje rezultate (primećeno i za Undersampling i za Oversampling) pušten je Random Forest klasifikator kada su se podaci podelili na tačno 50-50% pripadnosti u obe klase. Izračunato je da se 209 instanci poveća za 2148 procenta da bi došlo do istog broja.

Sada je dalo bolje rezultate

```
=== Summary ===

Correctly Classified Instances        8938               95.0952 %
Incorrectly Classified Instances       461                4.9048 %
Kappa statistic                          0.9019
Mean absolute error                      0.0945
Root mean squared error                  0.2023
Relative absolute error                 18.9049 %
Root relative squared error             40.4657 %
Total Number of Instances             9399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943    0.041    0.958      0.943    0.951      0.902    0.987     0.986     0
                 0.959    0.057    0.944      0.959    0.951      0.902    0.987     0.988     1
Weighted Avg.    0.951    0.049    0.951      0.951    0.951      0.902    0.987     0.987

=== Confusion Matrix ===

    a    b   <-- classified as
 4431  268 |    a = 0
  193 4507 |    b = 1
```



Correctly classified instances je 8938, a incorrectly je 461, dok je za prethodni oversample dao 8600 correctly a incorrectly 426

FILTERI s ovakvim oversample-ovanjem:

Attribute Selection Filters tj. filteri za izbor atributa: Ovi filteri imaju za cilj da izaberu podskup relevantnih karakteristika za klasifikaciju. Oni pomažu u smanjenju dimenzionalnosti i poboljšanju efikasnosti i tačnosti klasifikatora. Neki filteri za izbor atributa u Weki koje smo probali su InfoGainAttributeEval, ChiSkuaredAttributeEval i CfsSubsetEval

- CfsSubsetEval

Koji je izdvojio atribute s forward direction:

1 ☐ age
2 ☐ ever_married
3 ☐ Residence_type
4 ☐ stroke

Imao rezultate:

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

Correctly Classified Instances        8750               93.095  %
Incorrectly Classified Instances       649                6.905  %
Kappa statistic                          0.8619
Mean absolute error                      0.0924
Root mean squared error                  0.2254
Relative absolute error                 18.4824 %
Root relative squared error             45.0745 %
Total Number of Instances             9399

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.957     0.095     0.910       0.957   0.933      0.863    0.977     0.967     0
               0.905     0.043     0.954       0.905   0.929      0.863    0.977     0.981     1
Weighted Avg.  0.931     0.069     0.932       0.931   0.931      0.863    0.977     0.974

=== Confusion Matrix ===

    a    b   <-- classified as
 4495  204 |    a = 0
  445 4255 |    b = 1
```

Koji je bukvalno identične rezultate dao za bi-directional i backward.

- InfoGainAttributeEval,procenjuje vrednost atributa merenjem dobiti informacije u odnosu na klasu.

evaluator Choose **InfoGainAttributeEval**

search Choose **Ranker** -T -1.7976931348623157E3

```
=== Summary ===

Correctly Classified Instances        8923               94.9356 %
Incorrectly Classified Instances       476                5.0644 %
Kappa statistic                          0.8987
Mean absolute error                      0.0956
Root mean squared error                  0.2036
Relative absolute error                 19.127  %
Root relative squared error             40.7226 %
Total Number of Instances             9399

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.940     0.041     0.958       0.940   0.949      0.899    0.987     0.986     0
               0.959     0.060     0.941       0.959   0.950      0.899    0.987     0.988     1
Weighted Avg.  0.949     0.051     0.950       0.949   0.949      0.899    0.987     0.987

=== Confusion Matrix ===

    a    b   <-- classified as
 4418  281 |    a = 0
  195 4505 |    b = 1
```

Oversampling opet

```
=== Summary ===

Correctly Classified Instances        8971               95.4666 %
Incorrectly Classified Instances       426                4.5334 %
Kappa statistic                          0.9093
Mean absolute error                      0.0858
Root mean squared error                  0.1912
Relative absolute error                 17.1667 %
Root relative squared error             38.2376 %
Total Number of Instances             9397

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.949    0.040    0.960      0.949    0.954      0.909    0.989     0.987     0
                 0.960    0.051    0.949      0.960    0.955      0.909    0.989     0.990     1
Weighted Avg.    0.955    0.045    0.955      0.955    0.955      0.909    0.989     0.988

=== Confusion Matrix ===

    a    b   <-- classified as
 4459  240 |    a = 0
  186 4512 |    b = 1
```

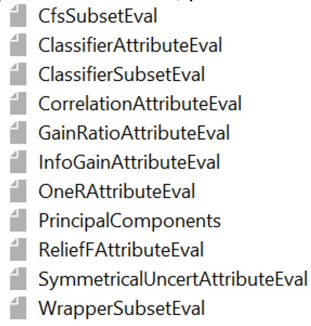
Testiranjeeeeeee

```
Dataset                   (1) trees.J | (2) baye (3) lazy (4) lazy (5) lazy (6) lazy (7) tree (8) tree (9) func (10
----------------------------------------------------------------------------------------------------------------------
jovonanovo-weka.filters.u  (0)        |
healthcare-dataset-stroke  (0)        |
healthcare-dataset-stroke(100)   0.93 |   0.82 *   0.94 v   0.94     0.93     0.94     0.96 v   0.92 *   0.83 *
----------------------------------------------------------------------------------------------------------------------
                            (v/ /*) |  (0/0/1)  (1/0/0)  (0/1/0)  (0/1/0)  (0/1/0)  (1/0/0)  (0/0/1)  (0/0 (0/0/0
```

Prethodno izbačene instance sa nepoznatim vrednostima za atribut 'bmi' su sada ušle u obzir prilikom klasifikacije, kod kojih su te nepoznate vrednosti stavljene na prosečnu vrednost tog atributa. Razmatrano je da se stavi na vrednost koja je medijana, međutim weka ne pruža tu opciju. Primećeno je da malo bolje vrednosti daje dataset kod kojih su izbačene instance koje nemaju vrednost za bmi, ali taj broj instanci je mali u odnosu na broj ukupnih instanci.

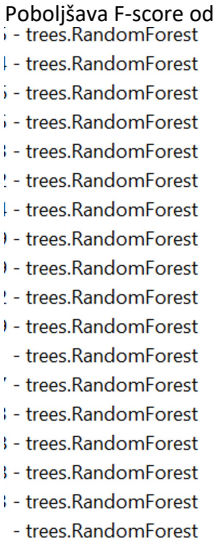
```
Dataset                   (1) trees.J | (2) baye (3) lazy (4) lazy (5) lazy (6) lazy (7) tree (8) tree (9) func (10
----------------------------------------------------------------------------------------------------------------------
healthcare-dataset-stroke(100)   0.93 |   0.82 *   0.94 v   0.94     0.93     0.94     0.96 v   0.92 *   0.83 *
jovonanovo-weka.filters.u(100)   0.92 |   0.80 *   0.94 v   0.93 v   0.93     0.93 v   0.95 v   0.92 *   0.83 *
----------------------------------------------------------------------------------------------------------------------
                            (v/ /*) |  (0/0/2)  (2/0/0)  (1/1/0)  (0/2/0)  (1/1/0)  (2/0/0)  (0/0/2)  (0/0 (0/0/0
```

Sada oversample-ovani skup podataka, kome je nepoznata vrednost za atribut 'bmi' stavljena na prosečnu vrednost, prolaze kroz razne filtere za selekciju atributa.



Probani su svi filteri i poneke kombinacije različitih filtera, ali samo jedan filter je dao boljitak u pogledu metrike, a taj filter je GainRatioAttributeEval-S.

Poboljšava F-score od 0.948 na 0.950.



Posle toga iskorišćen je filter Resample.
Filter Resample se koristi za ponovno uzorkovanje podataka radi rešavanja problema neravnoteže klase ili za kreiranje uravnoteženih skupova podataka za obuku modela mašinskog učenja. Ponovno uzorkovanje uključuje manipulisanje distribucijom klasa u skupu podataka bilo prekomernim uzorkovanjem manjinske klase, poduzorkovanjem većinske klase ili oboje.

"Resample" filter u Veki se koristi za modifikovanje distribucije instanci u skupu podataka. Može biti od pomoći u rešavanju problema neravnoteže klasa, gde je broj instanci u različitim klasama značajno neuravnotežen. Kada primenite "Resample" filter, prilagođava broj instanci u svakoj klasi da bi se postigla uravnoteženija distribucija. Filter pruža opcije za povećanje ili smanjenje broja instanci u klasi, kao i opcije za nasumično podešavanje redosleda instanci

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

Correctly Classified Instances        9489               97.6034 %
Incorrectly Classified Instances       233                2.3966 %
Kappa statistic                          0.9521
Mean absolute error                      0.0585
Root mean squared error                  0.1442
Relative absolute error                 11.7097 %
Root relative squared error             28.835  %
Total Number of Instances             9722

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.970    0.018    0.981      0.970    0.976      0.952    0.996     0.996     0
                 0.982    0.030    0.971      0.982    0.976      0.952    0.996     0.997     1
Weighted Avg.    0.976    0.024    0.976      0.976    0.976      0.952    0.996     0.996

=== Confusion Matrix ===

    a    b   <-- classified as
 4703  144 |    a = 0
   89 4786 |    b = 1
```

Normalizacija

Nakon što se normalizovale vrednosti, a to je bilo potrebno jer numeričke karakteristike u skupu podataka mogu imati različite razmere, opsege ili jedinice.
Na primer, jedna karakteristika može da se kreće od 0 do 100, dok druga karakteristika može da se kreće od 1000 do 100000. Takve razlike u razmerama mogu dovesti do netačnih rezultata tokom analize ili obuke modela. Normalizacija dovodi sve karakteristike u zajedničku skalu, obezbeđujući da podjednako doprinose analizi ili procesu učenja modela. Poboljšava performanse modela.

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

Correctly Classified Instances        9494               97.6548 %
Incorrectly Classified Instances       228                2.3452 %
Kappa statistic                          0.9531
Mean absolute error                      0.0588
Root mean squared error                  0.1443
Relative absolute error                 11.7603 %
Root relative squared error             28.8647 %
Total Number of Instances             9722

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.971    0.018    0.982      0.971    0.976      0.953    0.996     0.996     0
                 0.982    0.029    0.971      0.982    0.977      0.953    0.996     0.997     1
Weighted Avg.    0.977    0.023    0.977      0.977    0.977      0.953    0.996     0.996

=== Confusion Matrix ===

    a    b   <-- classified as
 4706  141 |    a = 0
   87 4788 |    b = 1
```

Diskretizacija

Diskretizacija je proces transformacije kontinuiranih numeričkih obeležja u diskretne ili kategoričke atribute. Uključuje podelu opsega vrednosti na intervale ili binove i dodeljivanje svake vrednosti određenom binu. The "Discretize" filter u

Weki nam omogućava da izvršimo ovu transformaciju. Kada primenimo opciju "Discretize" filter, odredimo broj binova ili intervala za kreiranje. Weka automatski određuje granice bina na osnovu distribucije vrednosti u atributu. Cilj je uhvatiti osnovne obrasce i varijacije u podacima uz smanjenje broja različitih vrednosti.

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

Correctly Classified Instances        9531               98.0354 %
Incorrectly Classified Instances       191                1.9646 %
Kappa statistic                          0.9607
Mean absolute error                      0.0449
Root mean squared error                  0.1301
Relative absolute error                  8.9784 %
Root relative squared error             26.0158 %
Total Number of Instances             9722

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Clas
                 0.983    0.023    0.977      0.983    0.980      0.961    0.997     0.996     0
                 0.977    0.017    0.983      0.977    0.980      0.961    0.997     0.997     1
Weighted Avg.    0.980    0.020    0.980      0.980    0.980      0.961    0.997     0.996

=== Confusion Matrix ===

    a    b   <-- classified as
 4767   80 |    a = 0
  111 4764 |    b = 1
```

Stavljen je broj binova na 10, binRangePrecision je 6.

* Osetljivost (sensitivity) – koliko dobro klasifikator prepoznaje pozitivne slogove

* Specifičnost (specificity) – koliko dobro klasifikator prepoznaje negativne slogove

* Preciznost (precision) – procenat pozitivno klasifikovanih slogova koji su zaista pozitivni

* Kompletnost (recall) – procenat ispravno klasifikovanih pozitivnih slogova u odnsou na ukupan broj pozitivnh slogova

* F mera (F measure, F1 score, F-score) - harmonic mean. Daje podjednak značaj preciznosti i kompletnosti.