Laboratorijska vežba 5 – Data Mining

Cilj vežbe: Upoznavanje sa algoritmima za klasifikaciju

U okviru ove vežbe koristiće se:

- Weka – Java biblioteka za mašinsko učenje,
- ace.arff – podaci za binarnu klasifikaciju,
- dud.arff – podaci za višeklasnu klasifikaciju,
- diabetes.arff – podaci za binarnu klasifikaciju,
- segment-challenge.arff i segment-test.arff – podaci za višeklasnu klasifikaciju

Instalacija potrebnih alata

- Preuzeti instalacioni fajl za Weka biblioteku sa sajta: W neka
- Pokrenuti preuzeti izvršni fajl i pratiti upustva za instalaciju

Weka terminologija

- Relacija = skup podataka (npr. skup označenih primera)
- Atribut = karakteristika objekta
- Instanca = konkretan primer koga u slučaju klasifikacije karakteriše niz atributa i oznaka klase
- Tip = odnosi se na tip atributa ili tip klase

Klasifikacija

U ovom delu vežbe biće opisan način kreiranja različitih modela za klasifikaciju korišćenjem Weka biblioteke. Za proces klasifikacije potrebno je učitati i pripremiti podatke, odabrati klasifikator i podesiti njegove parametre i naposletku, evaluirati klasifikaciju.

Domen problema 1

Potrebno je predvideti na osnovu zadatih karakteristika da li će se određene hemijske supstance (ligandi) vezati za proteine (klasa *aktivno* (došlo je do vezivanja) i *neaktivno*). Karakteristike su zadate u obliku 1024 bita, gde svaki bit označava prisustvo ili odsustvo određene osobine. 1025. atribut označava pripadnost klasi. U slučaju višeklasne klasifikacije, takođe postoji 1024 atributa koji se odnose na hemijske komponente, dok klasa označava za koji protein (od 40) će se supstanca vezati ili se neće vezati ni za jedan (41. klasa).

Korišćeni podaci

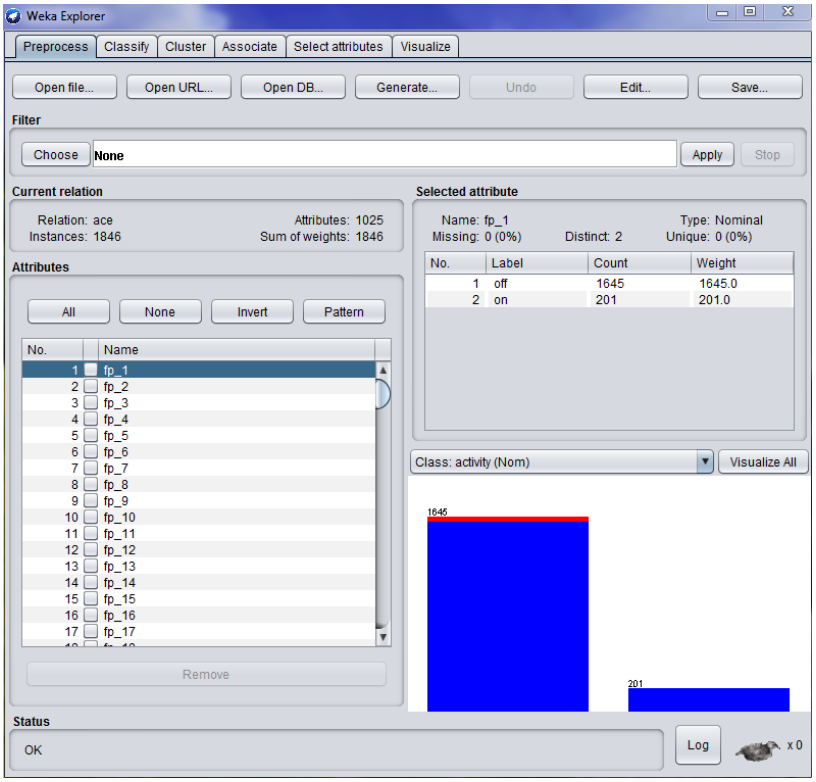
Podaci za binarnu klasifikaciju nalaze se u **ace.arff** fajlu. Postoji ukupno 1846 označenih primera. Za višeklasnu klasifikaciju, podaci se nalaze u **dud.arff** fajlu. U dud.arff fajlu postoji 7088 instanci i postoji 41 različita klasa.

ARFF (attribute-relation file format) je vrsta ASCII tekstualnog fajla koji opisuje niz primera sa odgovarajućim skupovima atributa. Više informacija je dostupno na linku: ARFF

Zadatak 1 – Naive Bayes

Binarna klasifikacija

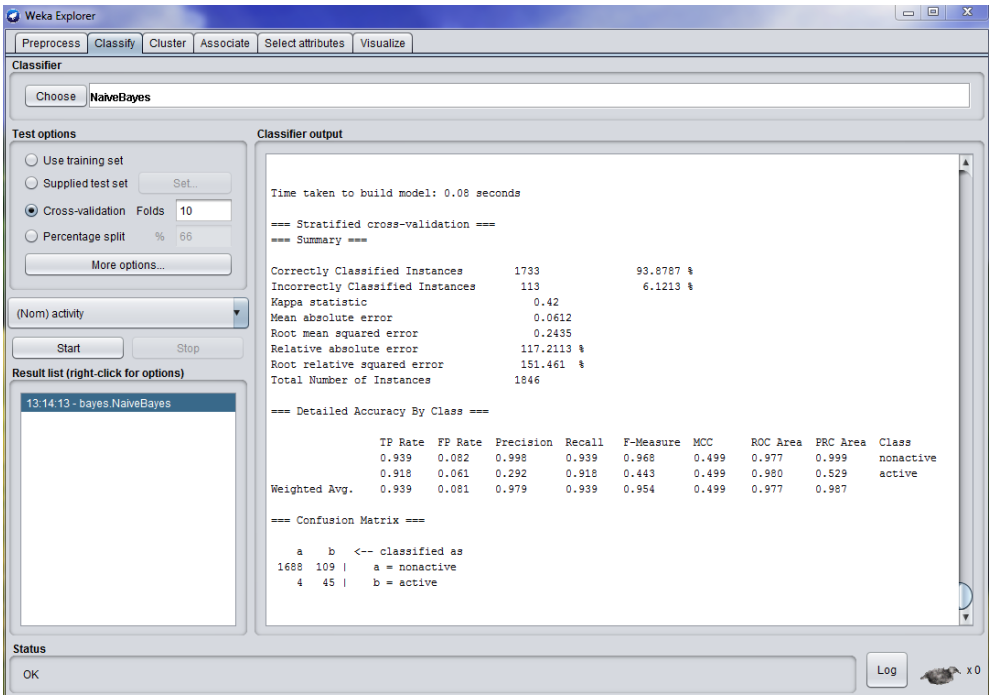
1. Pokrenuti Weku.
2. Po otvaranju grafičkog interfejsa, odabrati opciju **Explorer.**
3. Po otvaranju **Explorer** prozora, aktivna je **Preprocess** kartica za učitavanje i preprocesiranje podataka. Odabrati opciju **Open file** i izabrati ace.arff. Tada prozor izgleda kao na slici 1.
   U delu **Current relation** mogu se videti osnovne karakteristike izabranih podataka: naziv skupa podataka, broj atributa, broj instanci, suma težina instanci.
   U delu **Attributes** su prikazani svi atributi. Po potrebi, neki od atributa koji postoji u originalnom skupu se može ukloniti.
   U delu **Selected attributes** se prikazuju podaci o selektovanom atributu: koje su moguće vrednosti, koliko ima instanci sa svakom od mogućih vrednosti itd.
   Ukoliko se odabere poslednji atribut iz liste atributa (atribut *activity*), koji odgovara klasi, vidi se da postoje dve moguće vrednosti: aktivno i neaktivno. Ukupno postoji 1797 neaktivnih i 49 primera označenih klasom aktivno.



Slika 1 - Weka Explorer - učitavanje podataka

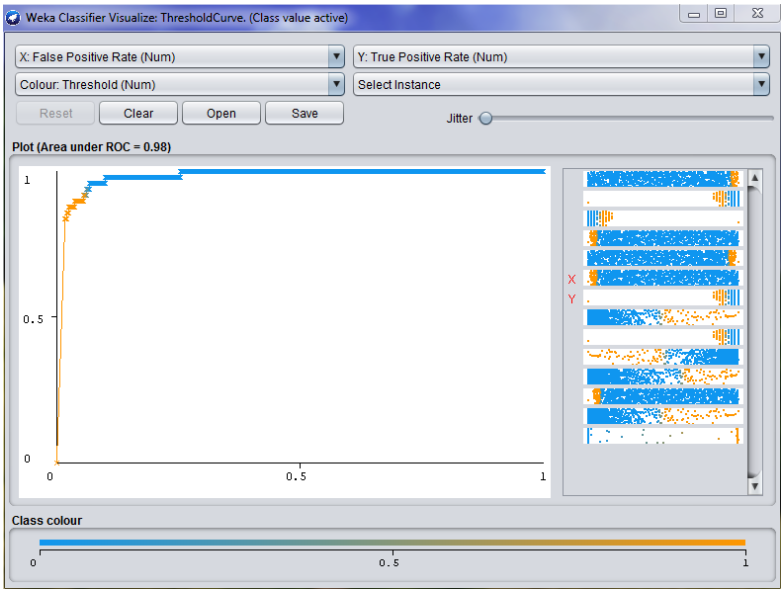
4. Odabrati karticu **Classify.**
5. U delu **Classifier,** odabrati opciju **Choose**. Ovom akcijom otvara se lista dostupnih algoritama.
6. Odabrati **weka/classifiers/bayes/NaiveBayes**
7. U delu **Test options** ostaviti podrazumevane vrednosti. Za testiranje se na taj način koristi cross-validation procedura.
8. Ispod opcija za testiranje, nalazi se padajući meni iz koga je moguće odabrati atribut koji će se koristiti kao oznaka klase. U ovom slučaju to treba da bude atribut *activity*.
9. Odabrati opciju **Start**.

10. Rezultati klasifikacije prikazani su na slici 2. U desnom delu Explorer prozora može se videti izlaz klasifikatora. Za konkretan primer postoji 1733 tačno klasifikovanih instanci i 113 pogrešno klasifikovanih, odakle sledi da je ukupna tačnost klasifikatora 93,88%. Pored ovih podataka prikazana je i Kappa statistika koja predstavlja meru korelacije između predviđenih i stvarnih klasa. Dodatno su prikazane i vrednosti različitih funkcija za procenu greške. Mogu se videti i sledeće mere: true positive rate, false positive rate, preciznost, odziv, F-mera, površina ispod ROC krive i druge, za svaku od klasa posebno.



Slika 2 - Rezultati klasifikacije - Naive Bayes

11. Vizuelizacijom rezultata je moguće dobiti korisne informacije o ponašanju klasifikatora. U delu **Result list**, desnim klikom odabrati opciju **bayes.NaiveBayes -> Visualize Threshold Curve -> active.** Rezultat ove akcije je prikazan na slici 3. U **Plot** delu prikazana je ROC kriva. Svaka tačka na ROC krivoj odgovara različitoj graničnoj vrednosti za podelu na dve klase, dok boja tačaka na krivoj odgovara predviđenoj vrednosti.



Slika 3 - ROC kriva - Naive Bayes

12. U **Weka Explorer** prozoru, na kartici **Classify**, u odeljku **Result list,** desni klik na bayes.NaiveBayes opciju otvara padajući meni iz koja je moguće odabrati opciju **Save Model**. Na taj način je moguće sačuvati .model fajl koji se posle može učitati u Java program i koristiti za klasifikaciju.

Višeklasna klasifiakcija

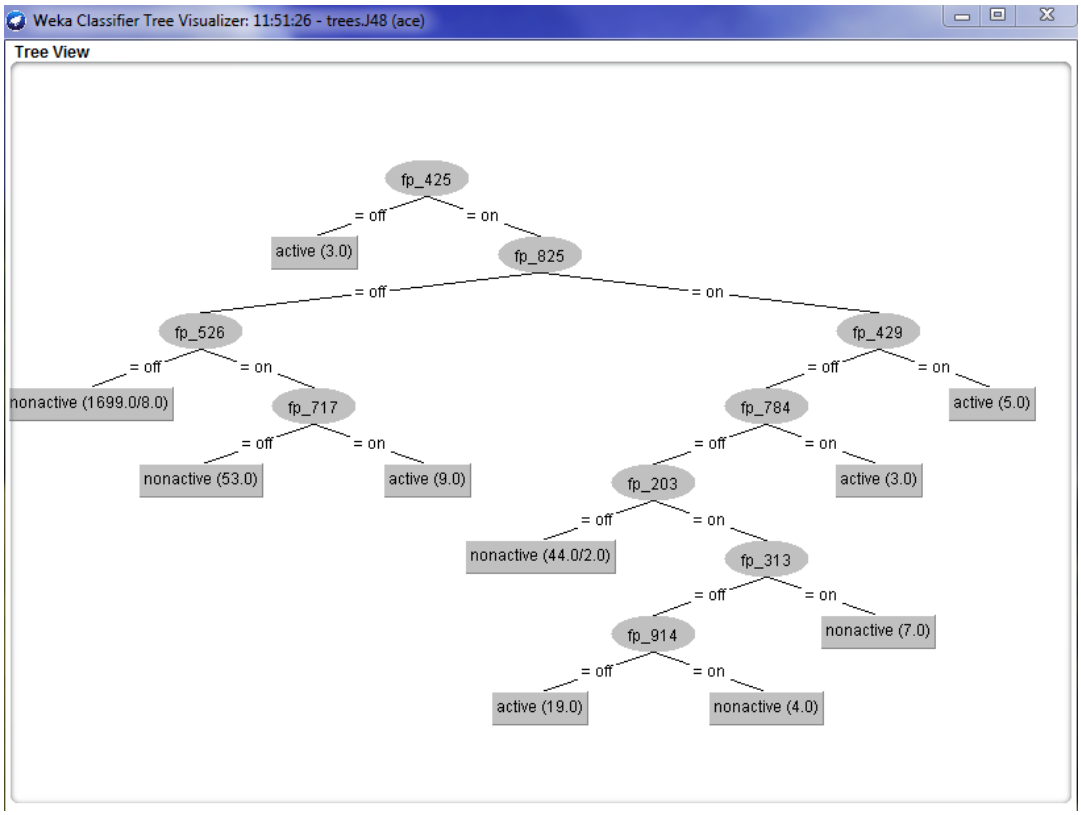
13. U **Weka Explorer** prozoru, vratiti se na karticu **Preprocess** i odabrati dud.arff fajl. Ukoliko se odabere *classes* atribut, u delu sa detaljima o selektovanom atributu može se videti da postoji 41 različita klasa, a raspodela instanci između klasa prikazana je na histogramu.
14. Preći na karticu **Classify**. Izabrati NaiveBayes, kao u koraku 6 ovog zadatka. Kliknuti na **Start** dugme za treniranje klasifikatora.
15. U **Classifier output** delu se može videti da tačnost klasifkatora iznosi približno 81% i da je korelacija 0.77, što ukazuje da model daje relevantno dobra predviđanja klasa.
16. Ispitati vrednosti različitih metrika algoritma za svaku od klasa. Može se videti da se različite klase predviđaju sa različitom tačnošću. Takođe se može videti i koji tip grešaka preovladava.

Zadatak 2 – Stabla odlučivanja

Binarna klasifikacija

1. U **Weka Explorer** prozoru, na kartici **Preprocess**, odabrati ace.arff fajl.
2. Na **Classify** kartici, odabrati **Choose->weka->classifiers->trees->J48**. Ostaviti podrazumevane vrednosti za ostala podešavanja: 10-fold cross-validation kao opciju za testiranje, *classes* kao atribut na osnovu koga će se vršiti klasfikacija. J48 je Weka implementacija C4 algoritma za kreiranje stabla odlučivanja.
3. Pregledati rezultate. Korelacija iznosi 0.76, što znači da je značajna.

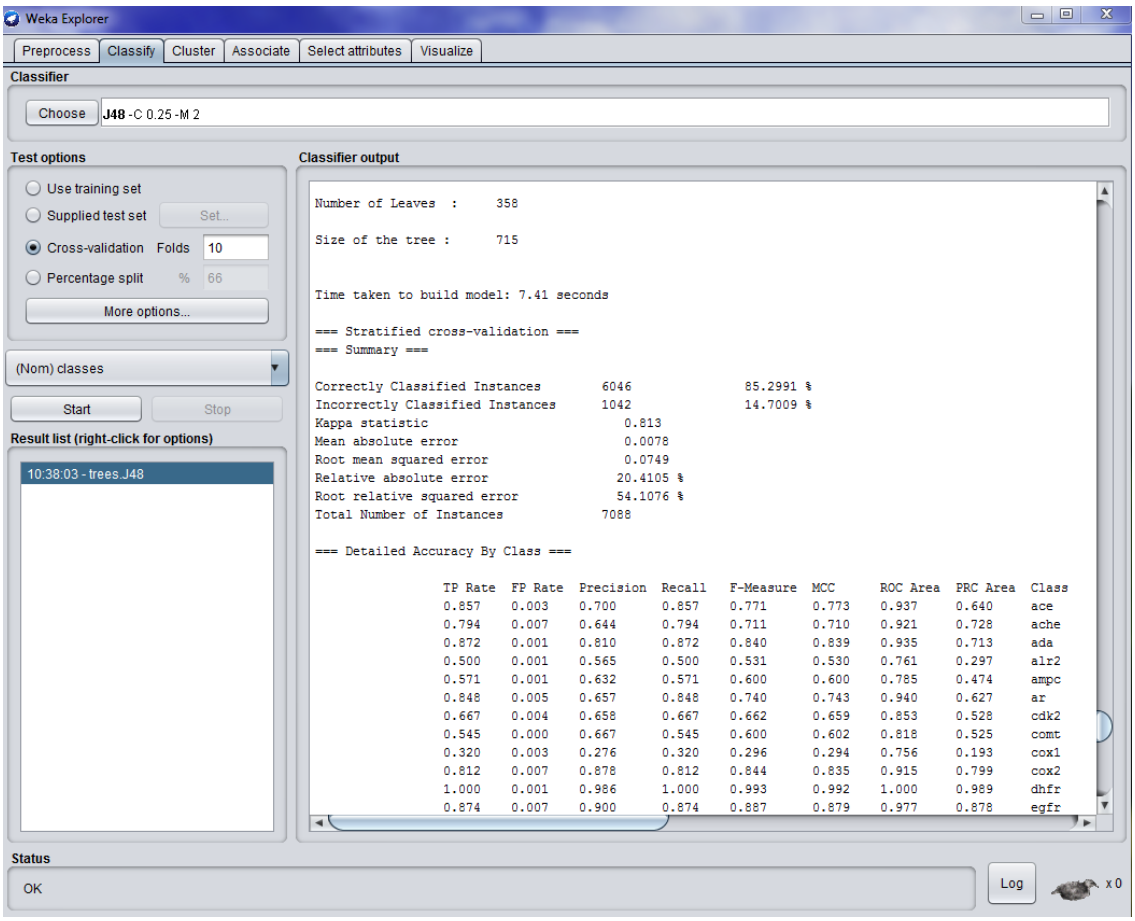
4. Kliknuti desnim klikom na opciju **trees.J48** u delu **Result list**, pa iz padajućeg menija odabrati opciju **Visualize tree**. U novom prozoru, kliknuti desnim klikom na grafičku reprezentaciju i iz padajućeg menija odabrati opciju **Fit to Screen**. Rezultat je prikazan na slici 5. Na ovaj način se mogu videti atributi koji se u svakom koraku koriste pri odlučivanju.



Slika 4 - Stablo odlučivanja za binarnu klasifikaciju

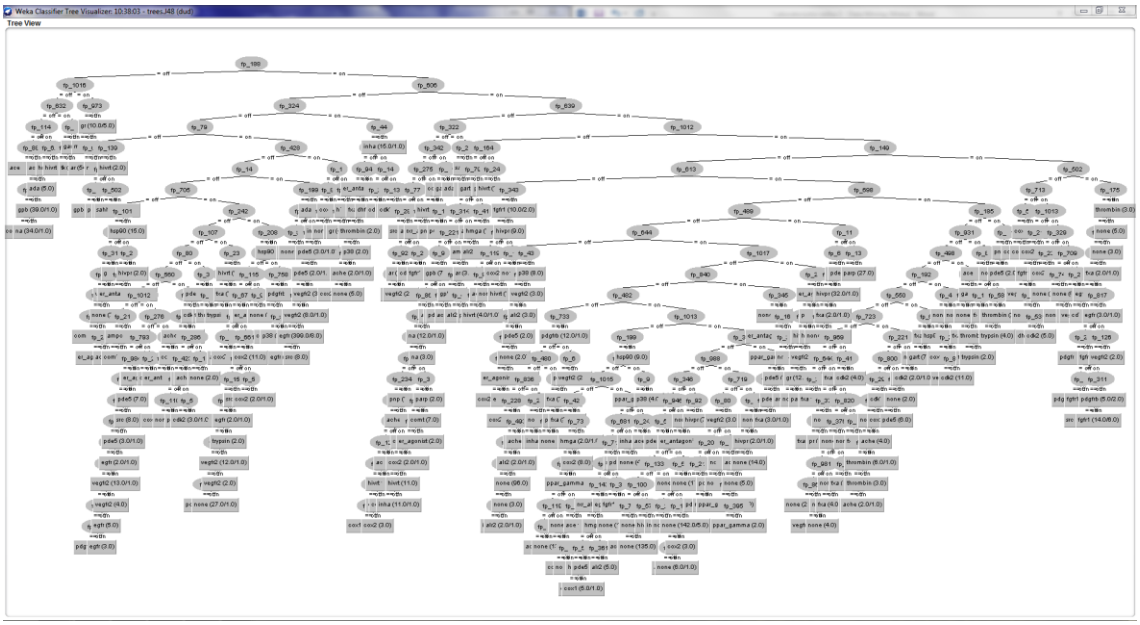
Višeklasna klasifikacija

1. U **Weka Explorer** prozoru, na kartici **Preprocess**, odabrati dud.arff fajl.
2. Na **Classify** kartici, odabrati **Choose->weka->classifiers->trees->J48**. Ostaviti podrazumevane vrednosti za ostala podešavanja: 10-fold cross-validation kao opciju za testiranje, *classes* kao atribut na osnovu koga će se vršiti klasfikacija.
3. Rezultat klasifikacije prikazan je na slici 6. Tačnost klasifikatora je približno 85%, što odgovara 6046 tačno klasifikovanih instanci. Takođe su za svaku klasu prikazane odgovarajuće mere – true positive rate, false positive rate, preciznost, odziv, F-mera itd. Treba zapaziti visoku vrednost Kappa statistic parametra – korelacija koja iznosi 0,813 značajno je veća nego kod Naive Bayes klasifikatora.



Slika 5 - Rezultat višeklasne klasifikacije stablom odlučivanja

4. Kliknuti desnim klikom na opciju **trees.J48** u delu **Result list**, pa iz padajućeg menija odabrati opciju **Visualize tree**. U novom prozoru, kliknuti desnim klikom na grafičku reprezentaciju i iz padajućeg menija odabrati opciju **Fit to Screen**. Rezultat je prikazan na slici 7. Na ovaj način se mogu videti atributi koji se u svakom koraku koriste pri odlučivanju.



Slika 6 - Vizualizacija stabla odlučivanja

Domen problema 2

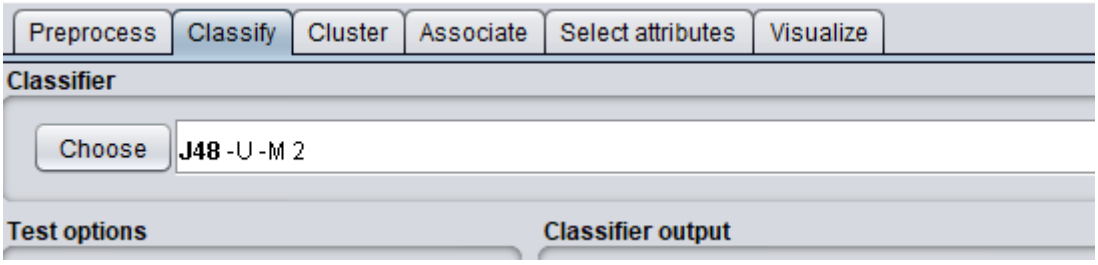
Potrebno je predvideti na osnovu zadatih atributa da li će rezultati testa na dijabetes biti pozitivni ili negativni. Set podataka je opisan sa ukupno 8 numeričnih atributa koje opisuju profil pacijanta, deveti atribut predstavlja klasu (0 – ako nema dijabetes, 1 – ako ima dijabetes).

Korišćeni podaci

Podaci se nalaze u **diabetes.arff** fajlu. Ukupno ima 768 instanci, od čega 500 pripada klasi 0 i 268 pripada klasi 1.

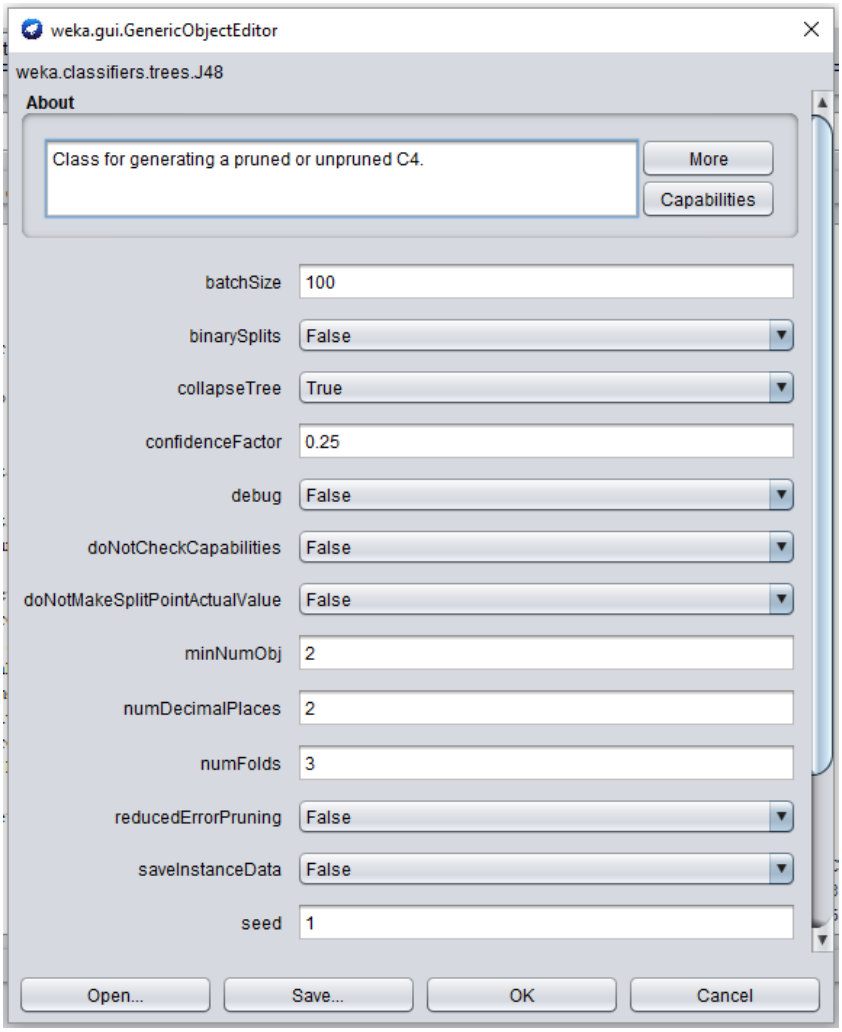
Zadatak 1 – J48 i J48 bez odsecanja

1. U **Weka Explorer** prozoru, na kartici **Preprocess**, odabrati diabetes.arff fajl.
2. Na **Classify** kartici, odabrati **Choose->weka->classifiers->trees->J48**.
3. Izabrati dodatna podešavanja algoritma klikom na ime algotirma (Slika 8).



Slika 7 - Podesavanja algoritma

4. Otvoriće se prozor prikazan an slici 9.



Slika 8 - Podešavanja J48 algoritma

5. Na dnu prozora nalazi se **unpruned** opcija koje podrazumevano ima vrednost FALSE. Ovo podešavanje označava proces smanjivanja stabla odlučivanja odstranjivanjem delova stable odlučivanja i podrazumevano je uključeno.
6. Bez promene ovog podešavanja kreirati stablo odluke za učitani set podataka.
7. Dobiveno stablo imaće 73.8% preciznost odlučivanja, ukupno 39 čvorova i 20 listova.
8. U sledećem koraku promeniti **unpruned** podešavanje na TRUE čime se isključuje smanjenje stabla odluke.
9. Sa ovakvim podešavanjem kreirati novo stablo odluke.
10. Novo stablo imaće preciznost 72.6% i ukupno 43 čvora i 22 lista.
    Primetićete da je stablo bez odsecanja veće ali i da ima manju preciznost pri odlučivanju. Nekada jednostavnije stablo odlučivanja daje bolje rezulate.

Zadatak 2 – Naive Bayes

1. Nad istim setom podataka izvršiti Naive Bayes algoritam.
2. Dobiveni razultati prikazani su na slici 10.

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

Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4664
Mean absolute error                      0.2841
Root mean squared error                  0.4168
Relative absolute error                 62.5028 %
Root relative squared error             87.4349 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.844     0.388     0.802      0.844     0.823      0.468    0.819     0.892     tested_negative
               0.612     0.156     0.678      0.612     0.643      0.468    0.819     0.671     tested_positive
Weighted Avg.  0.763     0.307     0.759      0.763     0.760      0.468    0.819     0.815

=== Confusion Matrix ===

   a   b   <-- classified as
 422  78 |   a = tested_negative
 104 164 |   b = tested_positive
```

Slika 9 – Rezultati

3. Primetićete da je pouzdnost ovog algoritma veća u odnosu na J48 stablo odlučivanja.
4. Pogledati ostale rezultate klasifikacije i ROC krivu.

Zadatak 3 – ZeroR

1. Nad istim setom podataka izvršiti ZeroR algoritam.
2. U **Weka Explorer** prozoru, na kartici **Preprocess**, odabrati diabetes.arff fajl.
3. Na **Classify** kartici, odabrati **Choose->weka->classifiers->rules->ZeroR**.
4. U delu **Test options** odabrati **Trening set**.
5. Odabrati opciju **Start**.
6. Dobiveni rezultati prikazani su na slici 11.

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

ZeroR predicts class value: tested_negative

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         158               68.6957 %
Incorrectly Classified Instances        72               31.3043 %
Kappa statistic                          0
Mean absolute error                      0.4495
Root mean squared error                  0.4666
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              230

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000     1.000     0.687      1.000     0.814      ?        0.500     0.687     tested_negative
               0.000     0.000     ?          0.000     ?          ?        0.500     0.313     tested_positive
Weighted Avg.  0.687     0.687     ?          0.687     ?          ?        0.500     0.570

=== Confusion Matrix ===

   a   b   <-- classified as
 158   0 |   a = tested_negative
  72   0 |   b = tested_positive
```

Slika 10 –Rezultati ZeroR klasifikatora

ZeroR je najjednostavniji klasifikator koji uvek bira onu klasi koja je bila u većini prilikom trenitanja podataka. U slučaju diabetes.arff podataka to je **tested_negative** klasa.

Ovaj klasifikator nad diabetes.arff podatacima ne daje dobre rezultate.

Zadatak 4 – OneR

7. Nad istim setom podataka izvršiti OneR algoritam.
8. U **Weka Explorer** prozoru, na kartici **Preprocess**, odabrati diabetes.arff fajl.
9. Na **Classify** kartici, odabrati **Choose->weka->classifiers->rules->OneR**.
10. U delu **Test options** odabrati **Percentage split** i upisati 70% za podelu podataka na trening i test set.
11. Odabrati opciju **Start**.
12. Dobiveni model klasifikatora prikazan je na slici 12.

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

plas:
        < 114.5 -> tested_negative
        < 115.5 -> tested_positive
        < 127.5 -> tested_negative
        < 128.5 -> tested_positive
        < 133.5 -> tested_negative
        < 135.5 -> tested_positive
        < 143.5 -> tested_negative
        < 152.5 -> tested_positive
        < 154.5 -> tested_negative
        >= 154.5        -> tested_positive
(587/768 instances correct)
```

Slika 11 – Primer modela OneR klasifikatora

OneR je jednostavan klasifikator koji bira jedan atribut i za njega definiše pravila na osnovu kojih se radi klasifikacija. U slučaju diabetes.arff podataka to je **plas** atribut.

Nakon izvšenog testiranja ovaj klasifikator je u 74.4% slučajeva uradio dobru klasifikaciju.

13. Pogledati ostale rezultate klasifikacije i ROC krivu.

Zadatak 5 – K Nearest Neighbors (kNN)

1. Nad istim setom podataka izvršiti OneR algoritam.
2. U **Weka Explorer** prozoru, na kartici **Preprocess**, odabrati diabetes.arff fajl.
3. Na **Classify** kartici, odabrati **Choose->weka->classifiers->lazy->IBk**.
4. U delu **Test options** cross valitation.
5. Odabrati opciju **Start**.
6. Dobiveni rezultati prikazani su na slici 13.

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

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

Correctly Classified Instances         539               70.1823 %
Incorrectly Classified Instances       229               29.8177 %
Kappa statistic                          0.3304
Mean absolute error                      0.2988
Root mean squared error                  0.5453
Relative absolute error                 65.7327 %
Root relative squared error            114.3977 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.794     0.470     0.759      0.794    0.776      0.331    0.650     0.732     tested_negative
               0.530     0.206     0.580      0.530    0.554      0.331    0.650     0.469     tested_positive
Weighted Avg.  0.702     0.378     0.696      0.702    0.698      0.331    0.650     0.640

=== Confusion Matrix ===

   a   b   <-- classified as
 397 103 |   a = tested_negative
 126 142 |   b = tested_positive
```

Slika 12 – Rezultati IBK klasifikatora

7. Analizirati rezultate i ROC krivu.
8. Klikom na ime klasifikatora otvoriti podešavanja. I promeniti vrednost polja KNN na 3.
9. Kliknuti na dugme **Start**.
10. Rezultati sa ovakvim parametrima prikazani su na slici 14.

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

Correctly Classified Instances         558               72.6563 %
Incorrectly Classified Instances       210               27.3438 %
Kappa statistic                          0.3822
Mean absolute error                      0.3092
Root mean squared error                  0.4525
Relative absolute error                 68.0324 %
Root relative squared error             94.9365 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.820     0.448     0.774      0.820    0.796      0.384    0.742     0.804     tested_negative
               0.552     0.180     0.622      0.552    0.585      0.384    0.742     0.569     tested_positive
Weighted Avg.  0.727     0.354     0.721      0.727    0.722      0.384    0.742     0.722

=== Confusion Matrix ===

   a   b   <-- classified as
 410  90 |   a = tested_negative
 120 148 |   b = tested_positive
```

Slika 13 - Rezultati IBK klasifikatora za K=3

Domen problema 3

U okviru ovog dela radi se analiza podataka koji predstavljaju karakteristike slika. Svaka instanca je opisana sa 19 atributa koje bliže opisuju deo slike, kao što su intenziteti boja i slično. Poslenji atribut je klasa koja opisuje šta je prikazano na slici i može uzeti jednu od sledećih vrednosti: brickface, sky, foliage, cement, window, path, grass.

Korišćeni podaci

Podaci za treninanje se nalaze u **segment-challenge.arff** fajlu. Za testiranje biće korišćen **segment-test.arff** fajl.

Zadatak

Isprobati rešavanje problema višeklasne klasifikacije sa sledećim klasifikatorima:

- J48
- K Nearest Neighbors (kNN) – za k=1, k=3 i k=5
- Naive Bayes

Uporediti tačnost, preciznost, odziv i korelaciju.