Zadaća br. 1

Datum objave:	**15.11.2021. god.**
Datum predaje zadaće:	**28.11.2021. god. u 23:59:59 h**
Ukupan broj bodova:	**10 bodova**

Za izradu zadaće se koristi *R Studio* okruženje i *R markdown* (Rmd) format koji se koristi na laboratorijskim vježbama. Izrada zadaće sastoji se od programskog koda (Rmd/HTML/PDF format) i dokumentacije (PDF format) sa izradom zadataka. Predaja zadaće se vrši preko Zamgera u nekom od podržanih formata za kompresovane *file*-ove (RAR, ZIP i sl.).

Zadaća se radi u timu. Svaki tim se sastoji od tri studenta pri čemu su sastavi timova određeni u koordinaciji sa predmetnim asistentom. Studenti koji još uvijek nisu raspoređeni u timove trebaju se javiti predmetnom asistentu putem emaila ekrupalija1@etf.unsa.ba, a najkasnije do 17.11.2021. god. Dovoljno je da samo jedan od studenata u timu izvrši predaju zadaće putem Zamgera.

Prije početka izrade zadatka 1 detaljno pročitati zadaću jer su *taskovi* međusobno povezani. Detaljno se upoznati sa predavanjima 1, 2, 3, 4 i 5 i vježbama 2, 3, 4 i 5. Na predavanjima 16.11.2021. god. će se analizirati i diskutovati tekst zadaće. Sva dodatna pitanja nakon diskusije na predavanjima za zadaću 1 postavljaju se putem foruma za zadaću 1 koji se nalazi na c2 najkasnije do 20.11.2021. god. u 23:59 h. Svi odgovori na pitanja biti će postavljeni u odgovarajućem *threadu* na forumu.

Cilj zadaće je izgradnja i evaluacija sljedećeg klasifikacijskog modela:

1. *Klasifikacijski model 1* (nastavne grupe 1, 2 i 3). Potrebno je, na osnovu seta podataka, izgraditi klasifikacijski model drveta odlučivanja koji će utvrditi da li će mušterija kompanije mrežnih usluga otkazati pretplatu. U tu svrhu koristiti će se labelirani (labela klase `Churn`) set podataka **customer_data_train.csv**.

2. *Klasifikacijski model 2* (nastavne grupe 4, 5 i 6). Potrebno je, na osnovu seta podataka, izgraditi klasifikacijski model drveta odlučivanja koji će utvrditi da li će na određenoj lokaciji sutra padati kiša. U tu svrhu koristiti će se labelirani (labela klase `RainTomorrow`) set podataka **weather_data_train.csv**.

Zadatak 1. (Istraživanje podataka) **(Ukupno: 3 boda)**

Upoznati se sa setom podataka i analizirati sve njegove karakteristike koristeći:

- osnovne metode deskriptivne statistike, kao i:
- metode procjene lokacije i varijabilnosti podataka;
- metode za procjenu korelacije između varijabli.

Na osnovu izvršene analize potrebno je preprocesirati podatke za primjenu drveta odlučivanja (pročitati i povezati sa zadatkom 2).

Preprocesiranje podataka treba između ostalog da uključi:

- popunjavanje svih nedostajućih vrijednosti;
- odbacivanje svih pronađenih outliera;
- odbacivanje atributa sa visokim stepenom korelacije;
- vršenje transformacija podataka.

Sve izvršene korake potrebno je detaljno dokumentovati i objasniti razloge za primjenu odabranih metoda za sve atribute seta podataka. Obavezno koristiti *ggplot2* napredne vizualizacijske funkcije za neke od grafika, na način kako je to urađeno u vježbi 2.

Zadatak 2. (Izgradnja modela klasifikacije) **(Ukupno: 5 bodova)**

a) Izgraditi sljedeće modele klasifikacije:

- drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit;
- drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks;
- C5.0 model klasifikacije.

Izgradnju modela vršiti na sljedeći način:

Prije izgradnje predikcijskih modela, na osnovu analize podataka potrebno je izvršiti sve potrebne korake preprocesiranja podataka (zadatak 1). Ovaj proces je potrebno dokumentovati na način da se prikažu poduzeti koraci, kao i da se daju objašnjenja i razlozi poduzimanja tih koraka. Zatim je potrebno izgraditi navedene klasifikacijske modele, dokumentovati i proces treniranja i testiranja modela, kao i rezultate evaluacije. Izvršiti čišćenje stabla za one modele za koje je to moguće. Ovaj korak se radi iterativno dok se ne dobiju rezultati za koje smatrate da su prihvatljivi.

Evaluacija istreniranih klasifikatora se vrši koristeći sljedeće metrike:

- konfuzijska matrica i mjere procjene koje se izvode na osnovu konfuzijske matrice (tačnost, osjetljivost, opoziv, F-mjera, itd.);
- ROC kriva;
- kappa statistika.

b) Implementirati predikcijske modele pod a) sa metodama:

- *holdouta* sa različitim podjelama na trening i testni podskup podataka;
- *k-fold* unakrsne validacije sa različitim vrijednostima parametra *k*;
- *k-fold bootstrapinga* sa različitim vrijednostima parametra *k*.

Analizirati koja je metoda je dala najbolja poboljšanja performansi klasifikatora.

c) Analizirati balansiranost podataka. Ukoliko set podataka nije balansiran, obavezno primijeniti neku od metoda za balansiranje seta podataka (*oversampling* ili *undersampling*) i izgraditi predikcijske modele iskazane pod a). Analizirati koja metoda je dala najbolja poboljšanja performansi klasifikatora.

d) Izgraditi sljedeće modele korištenjem ansambl tehnika za unaprjeđenje tačnosti klasifikacije:

- *bagging* model;
- *boosting* model (uz korištenje *AdaBoost* metode);
- *random forest* model.

Sve izvršene korake pod a), b), c) i d) potrebno je detaljno dokumentovati i izvršiti analizu i upoređivanje svih dobivenih rezultata. Obavezno koristiti različite načine za vizualizaciju konfuzijske matrice, na način kako je to urađeno u vježbi 4.

Zadatak 3. (Testiranje najboljeg modela) **(Ukupno: 1 bod)**

Odabrati najbolji od svih kreiranih modela klasifikacije iz zadatka 2 i izvršiti njegovu evaluaciju nad testnim podskupom podataka za odgovarajuću nastavnu grupu:

- za grupe 1, 2 i 3 koristiti će se set podataka **customer_data_test.csv**;
- za grupe 4, 5 i 6 koristiti će se set podataka **weather_data_test.csv**.

Dokumentovati dobivene rezultate. Da bi ovaj zadatak bio bodovan, potrebno je ispuniti sljedeće uslove:

- Tačnost veća od 0.75
- F1-mjera veća od 0.65
- Kappa statistika veća od 0.25
- Osjetljivost veća od 0.70
- Specifičnost veća od 0.75

Ukoliko odabrani model ne ispunjava sve uslove, neophodno je ponoviti i dokumentovati proces iz zadataka 1 i 2 dok se ne dobije model koji ispunjava tražene uslove. Bodovanje ove aktivnosti je u okviru bodova za zadatak 2.

Napomena: Zadatak će biti bodovan koristeći rangiranje svih timova koji rade na istom setu podataka. Svi timovi koji dobiju rezultate koji su bolji od minimalnih uslova iznad, biti će bodovani sa nagradnim bodovima. Maksimalan broj nagradnih bodova je 2. Pritom će prvoplasirani tim na rang-listi dobiti maksimalan broj nagradnih bodova, a svaki sljedeći tim 0.5 bodova manje od prethodnog tima na rang-listi.

Zadatak 4. (Pravilo-bazirana klasifikacija) **(Ukupno: 1 bod)**

Izgraditi sljedeće pravilo-bazirane modele klasifikacije:

- 1R model odlučivanja;
- RIPPER model odlučivanja.

Izvršiti treniranje i evaluaciju ovih modela na isti način kao u zadatku 2, a zatim i njihovu evaluaciju nad testnim podskupom podataka na isti način kao u zadatku 3.

Dokumentovati dobivene rezultate na isti način kao u prethodnim zadacima. Da bi ovaj zadatak bio bodovan, potrebno je ispuniti sljedeće uslove:

- Tačnost veća od 0.70
- Obuhvat veći od 0.80

Prikazati skup pravila koji dovodi do ispunjenosti ova dva uslova.