

Istraživanje podataka 2

Uputstvo za izradu projekta

Opšte informacije

Praktični deo ispita na kursu „Istraživanje podataka 2“ može se položiti na dva načina:

- Izlaskom na ispit u jednom od ispitnih rokova.
- Izradom timskog projekta tokom vežbi.

Ako se student odluči za polaganje praktičnog dela ispita putem projekta, potrebno je da formira odgovarajući tim. Svaki tim mora imati najmanje 2 člana, a može imati najviše 4 člana.

Timovi samostalno predlažu skup podataka nad kojim će raditi tokom semestra.

Preduslov za prijavljivanje projekta

Svaki član tima mora da otvori *GitLab* nalog, ukoliko ga već ne poseduje. Članovi tima moraju da se dogovore o skupu podataka na kojem će raditi projekat.

Prijava teme za projekat

Prijava se vrši popunjavanjem ankete koja će blagovremeno biti objavljena na stranici kursa, u delu sa obaveštenjima.

Samo jedan član tima prijavljuje projekat i unosi podatke za sve članove tima. Student koji prijavljuje projekat treba da navede sledeće informacije za svakog člana tima, uključujući i sebe:

- *GitLab* korisničko ime
- Broj indeksa
- Ime i prezime

Nakon isteka roka za prijavu projekata, asistent će otvoriti repozitorijume. Studenti ne treba samostalno da kreiraju repozitorijume za projekte.

Timovi rade isključivo u dodeljenom repozitorijumu, kako bi se mogao pratiti napredak projekta, kao i doprinos svakog člana tima.

Izbor skupa podataka

Skupove podataka možete potražiti na stranici [Kaggle](https://www.kaggle.com/).

Važno je da izabrani skup podataka bude dovoljno veliki, kako u pogledu broja instanci tako i u pogledu broja atributa. Minimalan broj instanci iznosi 2.000 instanci za studente I smera, odnosno 1.000 instanci za studente M smerova.

Broj atributa skupa podataka nije ograničen, ali se preporučuje¹ izbor skupa nad kojim je moguće primeniti i klasifikaciju i klasterovanje. Preporučuje se da prilikom izbora skupa podataka definišete ciljne *KPI* pokazatelje koje ćete pratiti tokom izrade projekta.

Očekivanja od projekta

Implementacija se vrši isključivo u programskom jeziku *Python*. Preporučuje se korišćenje alata *Jupyter Notebook* ili drugog sistema koji omogućava pisanje i izvršavanje koda, kao i dodavanje komentara i objašnjenja u istom fajlu.

Očekuje se da studenti primenjuju znanje stečeno na predavanjima i vežbama.

Iako je projekat timski, **svaki član tima** dužan je da primeni najmanje **jedan model klasterovanja i jedan model klasifikacije**. To znači da tim od 4 člana treba da primeni najmanje 8 različitih modela. Pored toga, **najmanje jedan od modela koje primeni svaki član tima mora biti model koji nije obrađivan na predmetu Istraživanje podataka 1 (IP1)**.

Za ostvarivanje maksimalnog broja poena, odnosno za dobijanje dodatnih poena, potrebno je primeniti i neki napredni model za obradu teksta, vremenskih serija, grafovskih podataka ili sličnih tipova podataka.

Ovakva podela posla omogućava svim članovima tima da paralelno rade na pretprocesiranju podataka i primeni modela. Na nivou celog tima potrebno je sprovesti kompletan proces *EDA* i *ETL*, kao i uporediti dobijene rezultate.

Očekivanja od izveštaja

U datoteci *README* potrebno je navesti izvor iz kojeg je skup podataka preuzet, kao i detaljno opisati skup nad kojim se radi.

Ako se koristi *Jupyter Notebook*, iznad svake ćelije sa kodom potrebno je odgovoriti na dva pitanja:

- Zašto je preduzet dati korak, odnosno zašto je pokrenut određeni model?
- Šta se dobija primenom tog koraka, odnosno koji su rezultati primene modela?

Na kraju projekta potrebno je uporediti i prokomentarisati rezultate koje su dobili svi članovi tima. Takođe, potrebno je utvrditi da li dobijeni rezultati potvrđuju prethodno definisane *KPI* pokazatelje ili se na osnovu sprovedene analize oni ne mogu potvrditi.

¹ U slučaju da nije moguće uraditi i klasifikaciju i klasterovanje nad istim skupom, moguće je imati i dva skupa. Ako se izaberu dva skupa podataka neophodno je da oba skupa prođu kroz *ETL* proces i pretprocesiranje podataka.