Istraživanje podataka - praktični deo ispit, jun 2018.

Broj indeksa	**Ime i prezime**

Zadaci se rade 120 minuta. Broj poena po zadacima je:

Zadatak	1	2	**Zbir**
maks	40	60	**100**
Osvojeno			

1. Na Desktopu u direktorijumu **ipJun2018** nalazi se skup podataka *DS_poslovi.xlsx* o traženim veštinama u oglasima koji se odnose na oblast Nauka o podacima. Primenom alata IBM SPSS Modeler i algoritma Apriori pronaći pravila pridruživanja o traženim znanjima. Postaviti uslove da je najmanja podrška za telo 2%, a pouzdanost pravila 50%.

   - U posebnoj tabeli izdvojiti za svaku instancu (tj. oglas) njene podatke i podatke o 2 najbolja pravila prema Lift meri. Izdvojiti pravila za koja važi da se sve stavke (iz tela i iz glave) pojavljuju u oglasu. Dozvoljeno je da oba pravila imaju istu glavu. Podatke iz tabele sačuvati u html datoteci.
   - Na osnovu dobijenog modela napraviti novi model koji sadrži samo pravila koja u glavi imaju *Python* i koja su zanimljiva po Lift meri.

   U komentarima odgovoriti na pitanja:

   - Koja pravila pridruživanja su najzanimljivija? Zašto?
   - Šta biste savetovali da uči osobi koja želi da napreduje, a koja već zna *Hive* i *SQL*? Zašto?
   - Kome ne biste savetovali da uči *Python*? Zašto?

   Radni tok eksportovati i dodeliti mu ime u formatu **SPSS_*vasBrojIndeksa***. Odgovore pišite u datoteku sa nazivom **SPSS_*vasBrojIndeksa_odgovori*.txt**.

2. Na Desktopu u direktorijumu **ipJun2018** nalazi se skup podataka *DS_poslovi_klasifikacija.csv* o traženim veštinama u oglasima koji se odnose na oblast Nauka o podacima. Koristeći alat KNIME izvršiti klasifikaciju nad skupom. Ciljni atribut je *statistic_software_required*.

   U radnom toku:

   - Napraviti model primenom drveta odlučivanja. Vrednosti atributa za pravljenje modela koristiti kao kategorije.
   - Promenom vrednosti za parametar k napraviti modele primenom k najbližih suseda. k uzima vrednost iz intervala [2, 157] sa korakom 2. Izdvojiti model koji ima najveću preciznost na test skupu.
   - Napraviti izveštaj (tabelu) koja sadrži odziv za svaku klasu iz test skupa u dobijenim modelima. U izveštaju izdvojiti podatke za model dobijen primenom drveta odlučivanja i za najbolji model prema ukupnoj preciznosti dobijen primenom k najbližih suseda. Tabela ima kolone *klasa, odziv, model* (moguće vrednosti su *tree* i *knn*).

- Matrice konfuzije sačuvati u csv datotekama. Imenom jasno označiti na koji model se odnose i koji skup.

U komentarima opisati dobijene modele i uporediti ih.

Radni tok eksportovati i dodeliti mu ime u formatu **KNIME_*vasBrojIndeksa***. Odgovore pišite u datoteci sa nazivom **KNIME_*vasBrojIndeksa_odgovori***.

Uputstvo za čuvanje rada: Na Desktopu napravite direktorijum sa nazivom u formatu **ip.Jun.2018.*ime.prezime.brojIndeksa*** gde umesto ime, prezime i broj indeksa stavite vaše podatke. Npr, `ip.Jun.2018.petar.petrovic.543_2014`. U tom direktorijumu čuvajte rešenja zadataka i datoteke sa odgovorima.