Istraživanje podataka - praktični deo ispita, jul 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 **ipJul2018** nalazi se skup podataka *zitarice_klasterovanje.xlsx* koji sadrži podatke o žitaricama. Koristeći skup i alat IBM SPSS Modeler izvršiti klasterovanje nad skupom primenom algoritama K-sredina i Kohonen. Pri pravljenju modela algoritmom K-sredina postaviti da se u najviše 35 iteracija izdvoji 8 klastera. Pri pravljenju modela algoritmom Kohonen menjati parametre. U komentarima odgovoriti na pitanja:

   - Da li je bilo obrade podataka pre klasterovanja? Zašto?
   - Koliki je kvalitet dobijenih modela?
   - Koje ste vrednosti zadavali za parametre pri pravljenju različitih modela primenom algoritma Kohonen? Ukratko napisati zaključak o dobijenim modelima.
   - Za svaki model zasebno (1 - dobijen algoritmom K-sredina, 2 - najbolji dobijeni model algoritmom Kohonen) uporediti najveći i najmanji klaster prema tri najznačajnija atributa za pravljenje modela. Koji su to atributi?

   Podatke o dobijenim modelima sačuvati u html datotekama.

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

2. Na Desktopu u direktorijumu **ipJul2018** nalazi se skup podataka *zitarice_klasifikacija.csv* koji sadrži podatke o žitaricama. Koristeći alat KNIME izvršiti klasifikaciju nad skupom. Ciljni atribut je *class*.

   U radnom toku:

   - Napraviti model primenom algoritma Naivni Bajes.
   - Promenom vrednosti za parametar k napraviti modele primenom k najbližih suseda. k uzima vrednost iz intervala [2, 10] sa korakom 1. Izdvojiti model koji ima najveću preciznost na test skupu.
   - Napraviti izveštaj (tabelu) koja sadrži preciznost za svaku klasu iz test skupa u dobijenim modelima. U izveštaju izdvojiti podatke za model dobijen primenom algoritma Naivni Bajes i za najbolji model prema ukupnoj preciznosti dobijen primenom algoritma k najbližih suseda. Tabela ima kolone *klasa, preciznost, model* (moguće vrednosti su *tree* i *knn*). Voditi računa da preciznost za neku klasu može biti i nedostajuća.
   - 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. Ukoliko je bilo obrade podataka pre primene klasifikacije, navesti šta i zašto ste radili.

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

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

Opis atributa skupa:

- *name*: ime žitarice
- *class*: klasa
- *calories*: kalorija
- *protein*: grama proteina
- *fat*: grama masti
- *sodium*: miligram natrijuma
- *fiber*: grama dijetetskih vlakana
- *carbo*: grama složenih ugljenih hidrata
- *sugars*: grama šećera
- *potass*: miligrami kalijuma
- *vitamins*: vitamini i minerali