# Istraživanje podataka 1 - pismeni deo ispita, jun 2020. g.

| **Broj indeksa** | **Ime i prezime** |
|---|---|
| | |

Zadaci se rade 150 minuta. Broj poena po zadacima je:

| **Zadatak** | 1 | 2 | 3 | 4 | 5 | **Zbir** |
|---|---|---|---|---|---|---|
| **maks** | 8 | 15 | 13 | 30 | 34 | **100** |
| *Osvojeno* | | | | | | |

1. Dati su podaci:

| Temperatura | Kašalj | Glavobolja | Umor |
|---|---|---|---|
| 36,8 | Da | Blaga | Ne |
| 38 | Ne | Srednjeg intenziteta | Da |
| 37,5 | Ne | Jaka | Da |
| 37,2 | Da | Ne | Ne |
| 39 | Da | Jaka | Da |
| 38,5 | Ne | Blaga | Ne |
| 39,2 | Da | Jaka | Da |
| 37,8 | Ne | Srednjeg intenziteta | Da |

   - Na dati skupu podataka se primenjuje algoritam koji koristi samo numeričke atribute. Kako biste transformisali dati skup podataka tako da je moguće primeniti taj algoritam i da se koriste informacije iz svih atributa? Ispisati transforimane prve dve instance zadatog skupa.
   - Nad transformisanim skupom se primenjuje algoritam koji koristi euklidsko rastojanje za poređenje dve instance. Da li je potrebno izvršiti preprocesiranje podataka? Obavezno obrazložiti odgovor.

2. Dati su podaci za treniranje:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **A** | a | a | a | b | b | a | b | b | a |
| **B** | s | f | s | s | s | f | f | f | s |
| **C** | m | m | m | m | m | n | n | n | n |
| **Klasa** | + | + | - | + | + | - | + | - | + |

   i podaci za testiranje:

| | | | | |
|---|---|---|---|---|
| **A** | a | b | b | a |
| **B** | s | s | f | f |
| **C** | n | n | m | m |
| **Klasa** | + | - | - | + |

Na osnovu trening podataka, korišćenjem naivnog Bajesovog algoritma, klasifikovati test podatke i izračunati preciznost.

3. Data je matrica rastojanja između instance A-E. Izvršiti hijerarhijsko klasterovanje korišćenjem *max* veze. Rezulatat prikazati dendogramom. Ukoliko je potrebno identifikovati tri klastera na ovaj način, koji bi to klasteri bili?

| | **A** | **B** | **C** | **D** | **E** |
|---|---|---|---|---|---|
| **A** | 0 | 1 | 2 | 3 | 5 |
| **B** | | 0 | 4 | 5 | 8 |
| **C** | | | 0 | 3 | 6 |
| **D** | | | | 0 | 4 |
| **E** | | | | | 0 |

4. Na Desktopu u direktorijumu **ipJun22020** nalazi se skup podataka *skup1.csv* sa podacima o kupovini u kafeu. Jedan red sadrži podatke: ime dana kada je izvršena kupovina, id transakcije i naziv kupljenog proizvoda. Primenom alata IBM SPSS Modeler i algoritma Apriori pronaći pravila pridruživanja.

   - Pronaći pravila pridruživanja o proizvodima koji se kupuju zajedno. Postaviti uslove da je najmanja podrška za telo 2%, a pouzdanost pravila 40%. Dobijeni model nazvati *model1*.
   - Koja pravila pridruživanja u modelu 1 su najzanimljivija? Zašto? Pravila na osnovu kojih se donosi zaključak sačuvati u html datoteku *PravilaGrupa1*.
   - Pronaći pravila pridruživanja o kupljenim proizvodima samo za transakcije koje su izvršene ponedeljkom i u kojima nije kupljena kafa. Postaviti uslove da je najmanja podrška za telo 2%, a pouzdanost pravila 35%. Dobijeni model nazvati *model2*.
   - Na osnovu modela 2 odgovoriti na pitanje: Koje pravilo je najzanimljivije prema Lift meri? Pravila koja su zanimljiva na osnovu Lift mere sačuvati u html datoteku *Model2Lift*.

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

5. Na Desktopu u direktorijumu **ipJun22020** nalazi se skup podataka *skup2.csv* sa podacima o različitim vrstama stakla. Izvršiti klasifikaciju nad tim skupom primenom drveta odlučivanja i K najbližih suseda i unakrsne validacije u programskom jeziku Python.

   - Primeniti stratifikovanu podelu na trening i test skup.
   - Napraviti različite modele klasifikacije primenom drveta odlučivanja i promenom vrednosti za bar tri parametra.
   - Za najbolji model prema preciznosti dobijen primenom drveta odlučivanja izdvojiti preciznost i matricu konfuzije za trening i test skup.
   - Za najbolji model prema preciznosti dobijen primenom drveta odlučivanja rezultat klasifikacije test skupa prikazati pomoću grafika sa razbacanim elementima (eng. scatter). Primeniti tehniku PCA radi smanjenja dimenzija skupa podataka na dva atributa i koristiti rezultat za grafički prikaz. Svakoj klasi dodeliti jedinstvenu boju.
   - Napraviti različite modele klasifikacije primenom algoritma K najbližih suseda i promenom vrednosti za bar dva parametra.

- Za najbolji model prema preciznosti dobijen primenom KNN izdvojiti preciznost i matricu konfuzije za trening i test skup.
- Za najbolji model prema preciznosti dobijen primenom KNN rezultat klasifikacije test skupa prikazati pomoću grafika sa razbacanim elementima (eng. scatter). Primeniti tehniku PCA radi smanjenja dimenzija skupa podataka na dva atributa i koristiti rezultat za grafički prikaz. Svakoj klasi dodeliti jedinstvenu boju.

U komentarima:

- Da li je bilo dodatne obrade podataka i zašto?
- Koji procenat varijanse je objašnjen sa prve dve komponente skupa dobijenog primenom PCA tehnike?
- Diskutovati dobijene modele i uporediti ih.

Skriptu dodeliti ime u formatu **klasifikacija_*vasBrojIndeksa***. Izlaz programa sačuvajte u datoteci sa nazivom u formatu **izlaz_*vasBrojIndeksa*.txt**. Odgovore pišite u datoteku sa nazivom **klasifikacija_*vasBrojIndeksa*_odgovori**.

**Uputstvo za čuvanje rada**: Na Desktopu napravite direktorijum sa nazivom u formatu **ip.Jun2.2020.*ime.prezime.brojIndeksa*** gde umesto ime, prezime i broj indeksa stavite Vaše podatke. Npr, `ip.Jun2.2020.petar.petrovic.543_2014` U tom direktorijumu čuvajte rešenja praktičnih zadataka i datoteke sa odgovorima.

**Opis atributa u skupu *skup2***

- *RI*: indeks loma
- *Na*: natrijum
- *Mg*: magnezijum
- *Al*: aluminijum
- *Si*: silikon
- *K*: kalijum
- *Ca*: kalcijum
- *Ba*: barijum
- *Fe*: gvožđe
- *Type*: ciljni atribut
  - *a*: zidani prozori obrađeni
  - **b**: zidani prozori neobrađeni
  - *c*: vozila
  - *d*: kontejneri
  - *e*: pribora za jelo
  - *f*: prednji farovi