Adatbányászati technikák
2. Házi feladat

Általános tanácsok a feladatokhoz:

- Először mindig olvassa végig a teljes feladatot, mielőtt nekilátna a megoldásnak!
- Az első két feladatot Weka Explorerben kell megoldania!
- Az első két feladat esetében az alpontoknál található kérdéseket kell megválaszolni, illetve helyenként képernyőképpel alátámasztani a feladatok megoldását.
- Ügyeljen arra, hogy a képernyőképek esetén a méret és a színek beállítása jól látható képet eredményezzen! **A túl kicsi, olvashatatlan képeket nem tudjuk értékelni**!
- Ahol magyarázatot kell adnia valamilyen jelenségre, ott a magyarázatot olyan részletességgel adja meg, hogy egy – az adatbányászat alapfogalmaival tisztában levő, de sem a konkrét adathalmazt sem az alkalmazott szoftvereket nem ismerő – laikus számára is könnyen érthető legyen.
- A harmadik feladatban Java kódot kell írnia, illetve egy rövid magyarázatot adnia a d) feladatban. A Java kódoláshoz az Eclipse IDE használata ajánlott, de nem kötelező.
- A feladatok szöveges megoldását (1-2 illetve 3/d feladatok) **pdf** formátumban kell beadni! A megoldások legyenek igényesen formázva!
- A harmadik feladat esetében a forrásfájlokat (.java kiterjesztésű fájlok) kell elküldeni! Amennyiben több forrásfájlt hoz létre, azokat közös mappába rakja! Ha a megoldását több különböző projektben készíti el, akkor ezek külön mappákba kerüljenek!
- A feladatok beadására e-mailben van lehetőség. A pdf dokumentumot valamint a forrásfájl(ok)at tartalmazó mappá(ka)t csomagolja egy zip állományba (**név_neptun.zip**), majd küldje el e-mailben a bagyibence@gmail.com címre, „**Adatbányászat HF név**” tárggyal (idézőjelek nélkül, a név helyére értelemszerűen a saját nevét kell írnia)!
- A beadási határidő azoknak, akik hivatalosan páratlan héten járnak gyakorlatra: **május 1. 11:59**
- A beadási határidő azoknak, akik hivatalosan páros héten járnak gyakorlatra: **május 8. 11:59**
- A feladatokat az utolsó laboralkalmon (13. illetve 14. hét) meg kell védeni. Ez az 1-2 feladatok esetében azt jelenti, hogy az adott magyarázatokat ki kell tudni fejteni, illetve kérésre az eredményeket reprodukálni kell tudni. A 3. feladat esetében a kód bizonyos részeinek megmagyarázására, illetve minimális módosítására kell készülni.
- A feladatok megoldásában az interneten található anyagok (Weka API, video tutorialok) rengeteg segítséget nyújthatnak.
- A feladat értelmezésével, és az esetleg felmerülő problémákkal kapcsolatban szintén a fenti e-mail címen tudnak segítséget kérni.

1. Feladat (osztályozás, 5 pont)

a) Töltse le az alábbi címen elérhető file-t! Ismerkedjen meg az adatokkal a file tetszőleges text editorban való megvizsgálásával! Mit jelentenek a célváltozó egyes értékei? (0,5 pont)

http://repository.seasr.org/Datasets/UCI/arff/mushroom.arff

b) Töltse be az adatokat a Weka Explorerbe! Melyik attribútum tartalmaz hiányzó értékeket? Az adatok hány százaléka hiányzik ennél az attribútumnál? Hagyjuk el ezt az attribútumot! (0,5 pont)

c) Mit gondol, a „stalk-shape", vagy az „odor" attribútum segítségével lehet jobban szétválasztani az eseteket? ***Adjon a válaszára vizuális bizonyítékot***, és ***magyarázza meg*** részletesen, hogy miből látható a válasza! (1 pont)

d) Alkalmazzon Naive Bayes osztályozót az adathalmazra! A tesztelési opciók közül azt válassza, amelyik a legmegbízhatóbb eredményt biztosítja! Írja le melyiket választotta, és röviden ***indokolja meg miért***! Végezze el az osztályozást, és adja meg a következő adatokat a teljes adathalmazra (1 pont):

- accuracy
- precision
- recall

e) Egyforma súlyú hiba-e a helytelen osztályozás mindkét lehetséges formája? ***Válaszát indokolja***! Ha nem, melyik a nagyobb probléma? Mennyi a hibás osztályozás költsége, ha a nagyobb hibát ötször akkorának tekintjük, mint a kisebbet (a kisebb hiba legyen egységnyi)? (1 pont)

f) Készítsen döntési fát, amely tökéletesen osztályozza az eseteket! A fában csak bináris elágazások lehetnek, és minden levélelemnek legalább 5 esetet kell tartalmaznia! A döntési fát elkészítő algoritmust és a többi paramétert szabadon megválaszthatja. ***Dokumentálja, hogy milyen osztályozót és paramétereket használt! A döntési fáról mellékeljen jól látható képet!*** (1 pont)

2. Feladat (klaszterezés, szabálykeresés, regresszió 5 pont)

a) Töltse le az alábbi URL-ről az adathalmazt! Az adatbázis a világ országairól és azok zászlóiról tartalmaz adatokat. Vizsgálja meg az adathalmazt! ***Milyen vallású Magyarország az adatbázis szerint?*** (0,5 pont)

http://repository.seasr.org/Datasets/UCI/arff/flags.arff

b) Az attribútumok közül hagyja el az ország nevét, területét és népességét tartalmazókat! Ezt követően végezzen klaszterezést az adatokon. A cluster mode ablaknál állítsa be, hogy a vallás attribútumot tekintse kiértékelési kritériumnak! Ezután végezzen Hierarchikus klaszterezést annyi klasztert megadva, ahány lehetséges értéke a vallás attribútumnak van. A távolságmértéket úgy adja meg, hogy két elem távolságának az egyes attribútumok értékkülönbségeinek összegét értsük. Állítsuk be azt is, hogy a klaszterek távolságának a két klaszter egymástól leginkább eltérő elemeinek távolságát tekintsük! ***Milyen előnyös tulajdonsága lesz az így kapott klasztereknek az elemszámot tekintve, más klaszterek közötti távolságértelmezésekkel szemben? Dokumentálja a*** leírás alapján választott ***paramétereket! Mellékeljen képet*** (vagy bemásolt logot) az így kapott klaszterezés confusion mátrixáról. (1,5 pont)

c) Végezzen klaszterezést K-means klaszterezővel is. A klaszterek számát és a távolságfüggvényt itt is a fentiekhez hasonlóan állítsa be! ***Mellékeljen képet*** (vagy logot) a klaszterezés kimenetéről! Ezzel a klaszterezővel ***jobb vagy rosszabb eredményt kaptunk*** mint a b) feladatban használt klaszterezővel? (0,5 pont)

d) Töltsük be újból az adathalmazt, ezúttal diszkretizáljuk a numerikus változókat a következő módon:

- A „rekeszek” száma legyen egyenlő a(z első) keresztnevében szereplő betűk számával
- Ha a vezetékneve magánhangzóra vagy kettős betűre végződik, akkor (nagyjából) azonos elemszámú intervallumokba, ha szimpla mássalhangzóra, akkor pedig egyforma szélességű intervallumokba sorolja az elemeket!

Dokumentálja az alkalmazott paraméter-beállításokat! Ezután végezzen apriori szabály-keresést, ahol olyan szabályokat keres, amelyek feltétel részét legalább 38 ország kielégíti, és legalább az esetek 85%-ában érvényes az adott szabály. ***Mellékeljen képet*** a beállításokról, és az 5 legerősebb szabályról! (1 pont)

e) Töltse be újból az adathalmazt, de ezúttal csak a terület, népesség, vallás és nyelv attribútumokat hagyja meg, a többit dobja el! Készítsen lineáris regressziót a területet függő változónak tekintve. Attribútum szelekciós módszernek válassza a mohó algoritmust, a többi beállítást hagyja változatlanul. A teljes adathalmazt tanító adatnak állítsa be. ***Milyen egyenletet kapott ezek alapján? Magyarázza el részletesen a regresszió jelentését*** (mit jelentenek a kapott együtthatók)! ***Melyik ország(ok) okozhatják elsősorban a nyelv és mely(ek) a vallás koefficiensének előjelét és mértékét? Mekkorára becsülnénk a nemrég megalakult Dél-Szudán területét, ha a lakosságot angol nyelvű (ez a hivatalos), keresztény (ez a leggyakoribb vallás), 12 millió fős országként határozzuk meg?*** (1,5 pont)

3. Feladat (Weka API használat, 5 pont)

a) Töltse le a következő URL címen található adatbázist egy tetszőleges mappába! Amennyiben szükségesnek tartja, olvassa el az adatbázisról szóló információkat a file elején.

http://repository.seasr.org/Datasets/UCI/arff/zoo.arff

b) ***A következő feladatok mindegyikében Java kódot kell írnia!*** Olvassa be a későbbi Weka használathoz megfelelő formában az adathalmazt, és állítsa be a célváltozót (az eredeti adatbázis „type" attribútuma). (0,5 pont)

c) Hozzon létre egy J48-as döntési fát, és állítsa be a következőket:

- A döntési fa legyen bináris!
- A levélelemek legalább 5 esetet tartalmazzanak!
- Kapcsolja be a „Reduced Error Pruning" funkciót!

Értékelje ki a létrehozott modellt 10-fold kereszt-validálással! ***Írassa ki a konzolra a modell teljesítményét összegző Stringet!*** (1,5 pont)

c) ***Írassa ki a konzolra a Confusion mátrixot*** úgy, hogy (a Weka explorerhez hasonló módon) az oszlopok és sorok fejlécében láthatók legyenek a célváltozó értékei! (1 pont)

d) A fenti adathalmazból szűrő segítségével hozzon létre egy új adathalmazt, amely nem tartalmazza a célváltozót! Ezt követően építsen SimpleKMeans klaszterező algoritmust az újonnan létrehozott adatbázisra (a k-t annyira állítsa, ahány lehetséges értéke a célváltozónak volt az eredeti adatbázisban). ***Az eredeti adatbázis segítségével értékelje ki***, hogy mennyire hatékonyan találta meg az algoritmus az eredeti osztályokat! Jobb, vagy rosszabb eredményt kaptunk mint a J48 fa esetében? Ez miből adódhat? (1 pont)

e) Törölje ki az adatbázisból az összes numerikus attribútumot! ***Figyelem, a célváltozó bár számokkal van kódolva, mégsem numerikus típusú!*** Hozzon létre asszociációs szabályokat az Apriori módszer segítségével! A minimális support alsó határát 0,3-ra, a minimális confidence értéket pedig 0,9-re állítsa! ***Írassa ki a konzolra az első 5 asszociációs szabályt!*** (1 pont)