**Alkalmazott statisztika**

**2. ZH**

**Gyakorlás**

**ADATBÁZISOK**

Az *admission.csv* adatbázis üzleti főiskolákra jelentkezők kétféle pontszámát tartalmazza (*GPA* és *GMAT*), valamint azt, hogy a jelentkező felvételt nyert-e (*De* = admit, notadmit, borderline), letölthető a következő webhelyről:

<http://www.biz.uiowa.edu/faculty/jledolter/DataMining/admission.csv>

A *car\_sales.csv* adatbázis különböző márkájú és típusú autók eladási adatait és egyéb tulajdonságait tartalmazza.

* manufact: márka
* model: típus
* sales: eladások száma (ezer)
* resale: viszonteladások száma (ezer)
* price: ár (ezer dollár)
* engine\_s: motorméret
* horsepow: lóerő
* wheelbas: tengelytáv
* width: szélesség
* length: hosszúság
* curb\_wgt: menetkész tömeg
* fuel\_cap: üzemanyagtartály űrtartalma
* mpg: üzemanyag-hatékonyság

A *vallalat.txt* adatbázis különböző vállalatok jellemzőit tartalmazza:

* NEV: a vállalat sorszáma
* MERET: a vállalat nagysága
* HATALMIT: hatalmi távolság
* KONZERVA: vállalat szemléletében jelen lévő konzervativizmus mértéke.

A *survey* adatbázis (*MASS* csomag) egyetemi hallgatók kérdőíves felmérésének eredményét tartalmazza:

* Smoke: dohányzási szokások
* Exer: a testmozgás gyakorisága

A *chickwts* adatbázis (*datasets* csomag) különböző fajta táplálékok csirkék tömegére vonatkozó hatását vizsgáló kísérletből származó adatokat tartalmaz:

* weight: a csirkék tömege
* feed: a táplálék típusa

Az *airquality* adatbázis (*datasets* csomag) napi levegőminőségi méréseket tartalmaz:

* Ozone: ózon mennyisége
* Temp: hőmérséklet

**1. feladat.**

Az *admission* adatbázisban található *GPA* és *GMAT* pontszámok alapján szeretnénk megjósolni a főiskolára való bekerülés valószínűségét (*De* = admit, notadmit, borderline).

Az alábbi utasítások segítségével olvassuk be az *admission.csv* adatbázist az *input* nevű táblázatba:

url ='http://www.biz.uiowa.edu/faculty/jledolter/DataMining/admission.csv'

input = read.csv(url)

1. Milyen módszert használ a felvételizők „admit”, „notadmit”, vagy „borderline” kategóriákba sorolására?
2. Térjünk át arra az új koordinátarendszerre, melyben a lehető legjobban szétválnak a csoportok. Adjuk meg a GPA változó első dimenzióhoz tartozó együtthatóját.
3. Adjuk meg az első megfigyelés *x*-koordinátáját ebben az új koordináta-rendszerben.
4. Az esetek hány százalékában működött jól az algoritmus?
5. Várhatóan melyik csoportba fog tartozni az a jelentkező, akinek a következő pontszámai vannak: GPA=3.21, GMAT=497? Mekkora valószínűséggel került ebbe a csoportba?

**2. feladat**

1. Olvassuk be a *car\_sales.csv* adatbázist.
2. Vizsgáljuk meg a folytonos változók közötti Pearson-korrelációkat! Mekkora a korreláció a lóerő (*horsepow*) és az ár (*price*) között?
3. Milyen módszer segítségével tudjuk az adatbázis méretét csökkenteni?
4. Hány korrelálatlan változóval jellemezhető az adatbázis, ha legfeljebb 15 százalékos információvesztést szeretnénk elérni? Pontosan mekkora az információvesztés mértéke?
5. Hogyan áll elő a legfontosabb új változó az eredetiek lineáris kombinációjaként? Adja meg az *mpg* változó együtthatóját.

**3. feladat**

1. Olvassuk be a *vallalat.txt* adatbázist.
2. Standardizáljuk a folytonos változókat.
3. Csoportosítsuk a vállalatokat a rendelkezésre álló folytonos változók alapján úgy, hogy jól szeparált klasztereket kapjunk. Milyen eljárást használ?
4. Melyik két vállalat került először egy csoportba, mekkora a köztük lévő távolság?
5. Mely csoportok lettek utoljára összevonva és mekkora a közöttük lévő távolság?
6. 2 csoport kialakításával mekkora lesz az egyes csoportok elemszáma és mely vállalatok tartoznak az egyes csoportokba?

**4. feladat**

1. Olvassuk be a „*MASS*” csomagból a „*survey*” adatbázist, mely egyetemi hallgatók kérdőíves felmérésének eredményét tartalmazza.
2. Teszteljük 5%-os szignifikancia szinten azt a nullhipotézist, hogy a dohányzási szokás (*Smoke*) független a testmozgás gyakoriságától (*Exer*). Megbízható-e a döntés? Miért?

**5. feladat**

1. Olvassuk be *chickwts* adatbázist.
2. Teszteljük 5%-os szignifikancia szinten azt a nullhipotézist, hogy a csirkék átlagos tömege (*weight*) független a táplálék típusától (*feed*).

**6. feladat**

1. Olvassuk be az *airquality* adatbázist.
2. Vizsgáljuk az ózon mennyisége (*Ozone*) és a hőmérséklet (*Temp*) közötti kapcsolatot.
   * Írjuk fel az ózonmennyiséget a hőmérséklet lineáris függvényeként. Adjuk meg az egyenes egyenletét. A hőmérséklet hány százalékban magyarázza meg az ózonmennyiséget?
   * Írjuk fel az ózonmennyiséget a hőmérséklet exponenciális függvényeként *Ozone* = exp(a + b\**Temp*) alakban. Adjuk meg az együtthatók becsléseit.
   * Melyik modell biztosítja a jobb illeszkedést?
   * Várhatóan mekkora lesz az ózonmennyiség 80 °F hőmérséklet esetén?