---
output:
  pdf_document: default
  html_document: default

Zárthelyi dolgozat pótlás

2022. május 6. Péntek

- Az érdemjegy az összes feladat elvégzésének eredményeképpen kerül meghatározásra.
- Maximális pontszám 100 pont.
- A zárthelyi megoldására 90 perc áll rendelkezésre.
- A félévközi aláírás megszerzéséhez legalább a 2-es (elégséges) szint elérése szükséges, ezen felül a kapott pontszám nem befolyásolja sem az aláírás megszerzését, sem a vizsgajegy meghatározását.
- A legutolsó mentés időpontja nem lehet későbbi, mint a dolgozat megkezdése +90 perc!!!
- Együttműködés gyanúja esetében az elért eredményt szóban fogjuk ellenőrizni!

91 -   5 (jeles)
81 - 90   4 (jó)
61 - 80   3 (közepes)
41 - 60   2 (elégséges)
0 - 40   1 (elégtelen)

1. R alapok (20 pont)

a) Készítsen négy, egyenként 20 elemű vektort a következők szerint. (5 pont)

- amount vektor: Függvénnyel 1 és 10 között egyenletesen generált véletlen egész számok.
- weight vektor: Függvénnyel normális eloszlás szerint 10 várható értékkel és 3 szórással generált véletlen valós számok. Figyeljen, hogy egyetlen szám sem legyeen 5 alatti (tehát az 5 alatti értékeket állítsa vissza 5-ös értékűre.)
- category vektor: Függvénnyel véletlenszerűen válaszva a "fruit" és a "vegetable" értékek közül
- color vektor: Függvénnyel véletlenszerűen választva a "red" és a "green" értékek közül

```{r}

```

b) Az a) feladatban elkészített vektorok segítségével hozzon létre egy árukészletet tartalmazó adatkeretet (stock), mely oszlop szerint tartalmazza az elérhető darabszámot (amount) az áru súlyát (weight), valamint az áru kategóriáját (category) és az termék színét (color). A category és color értékeket konvertálja faktorokká, valamint a sorokat nevezze el a sorszám szerint. (5 pont)

```{r}

```

c) Az adatkeret két kategorikus értéke (color és category), mivel kétállapotúak, összesen négy csoportot definiálhatunk. Az aggregate() függvény segítsével a négy csoportra (melyet a color és category különböző értékei határoznak meg) határozza meg az átlagos mennyiséget és az átlagos súlyt! (7 pont)

```{r}

```

d) Mentse el az adatkeretet csv fájlformátumban a sorok és oszlopok elnevezésével együtt. (3 pont)

```{r}

```

2. Vizualizáció (15 pont)

Ábrázolja a következő vonaldiagrammokat egy ábrában a -2 és 2 értéktartományon, 0.1 lépésközökkel.

- az x^2 függvényt, kék folytonos vonal segítségével, (3 pont)
- a 3 meredekségű egyenest, mely 2-nél metszi az y tengelyt, RGB=FF0011 színű folytonos vonallal, (3 pont)
- a sin(4x) + 2 függvényt, a curve() függvény segítségével, szaggatott zöld vonallal, (5 pont)
- lássa el az ábrát jelmagyarázattal (4 pont)

```{r}

```

3. Adatok megismerése (20 pont)

a) Töltse be az iris adathalmazt és szűrje ki a versicolor fajta Sepal Lenght értékeit. A vizsgálódásait ezeken az értékeken folytatja majd! (2 pont)

```{r}

```

b) Az adatsorból függvénnyel véletlenszerűen választva három adatot állítson át NA értékre! (3 pont)

```{r}

```

b) Határozza meg a kiszűrt adatsor centrumra és szórásra jellemző statisztikai jellemzőit! Természetesen az érvényes adatokra vonatkozóan! (10 pont)

```{r}

```

d) Tetszőlegesen választott módszer segítségével végezze el az adatok normalitásvizsgálatát! Az eredmény alapján szövegesen magyarázza! (5 pont)

```{r}

```

```{block id=szöveges válasz helye}

```

4. Regresszió - Predikció (20 pont)

a) Az iris adathalmaz setosa fajta esetében válassza ki azt a két (nem azonos) paramétert, melyek között legnagyobb a lineáris összefüggés! Választását indokolja! (5 pont)

```{r}

```

```{block}

```

b) A kiválasztott két paraméter közötti összefüggésre hozzon létre lineáris regressziós modellt. Az eredeti értékeket és a regresszió eredményét ábrázolja! (5 pont)

```{r}

```

c) A regresszió hibája (residual) alapján mit tud elmondani a modellalkotásról? (7 pont)

```{r}

```
```{block}

```

d) Mennyi lenne a függő változó értéke a lineáris regressziós modell alapján, ha a független változó értéke 4.5 lenne? (3 pont)

```{r}

```

5. Osztályozás (25 pont)
Az osztályozási feladat a `scat` adathalmazon (megtalálható a `caret` csomagban) történik. Az adathalmaz három állatfajta (bobcat,coyote,gray_fox) székletadatait tartalmazza.

a) Töltsük be az adathalmazt!

```{r}

```

b) Készítsünk egy adathalmazt, amely csak a `Species,Site,Location,Length,Diameter,Mass` változókat tartalmazza! Amennyiben az adathalmaz tartalmaz `NA` értékeket, távolítsuk el azokat a sorokat, amelyekben megtalálhatók! (5 pont)

```{r}

```

c) Válasszuk szét az adathalmazt teszt és tanuló adathalmazra! A tanuló adathalmaz az adatok 70%-a legyen! Figyeljünk oda, hogy osztályarányosan végezzük a szétválasztást a `Species` faktor alapján! (5 pont)

```{r}

```

d) Készítsünk egy döntési fa osztályozást az adathalmazra! Ehhez használjuk célváltozónak a fajt (`Species`), és az összes többi változót jellemzőnek! Keressük meg a fa optimális mélységét boostrap validációval 1 és 5 mélység között! Melyik lett az optimális mélység? (4 pont)

```{r}

```

e) Ábrázolja a döntési fát! Melyik változó volt a legfontosabb az osztályozás során? (3 pont)

```{r}

```

f) Készítsük el az osztályozáshoz tartozó keveredési mátrixot, mind a tanító, mind a teszt adathalmazhoz! Értékelje ki pár mondatban az osztályozás eredményét! (4 pont)

```{r}

```

g) Javítható az osztályozás, ha a tanító adathalmazt kiegyensúlyozzuk? Mutassa meg, hogy igen vagy sem (a minőség értékeléséhez használjuk a pontosság (accuracy) mutatót)! (4 pont)

```{r}

```