Félévközi feladatok

A feladat jellege egyszerű, egy független (az előadáson nem használt) adathalmazon való feladatmegoldást jelent, jellemzően a félév során elhangzott egy-egy témakör alapos megismerésén alapul. Számos feladat első része elméleti áttekintést is tartalmaz, ebben az esetben a gyakorlati feladat megoldást egy rövid **esszé** jellegű elméleti összefoglalás kell megelőzze az adott témakörről.

A mérnökileg precíz feladatmegoldás eredményét a feladatoknál megadott maximális pontszámmal jutalmazzuk. Maximális pontos az a feladatmegoldás, mely megfelelően meg van alapozva, a feladat megoldás helyes és körültekintő, a feladatmegoldás megfelelően meg van indokolva és az eredmények is jól alá vannak támasztva. Azaz lényegesen több indoklást várunk, mint egy ZH/vizsga feladat megoldásakor. A kirívóan precíz, körültekintő, vagy egy-egy témakörön túlmutató feladatmegoldásokat további plusz pontokkal jutalmazhatjuk.

Minden feladatot csak **egy hallgató** választhat! A félévközi feladatok **leadási határideje** az első vizsgaalkalom előtt két nap délután 5 óra. Azaz amennyiben az első vizsgaalkalomra való jelentkezés június 10-dikei dátummal történt, a feladat leadási határideje június 8, délután 5 (17:00) óra. Az elővizsgázók jellemzően az első hivatalos vizsgaalkalom előtt két nappal kell leadják a félévközi feladatmegoldásukat. A feladatmegoldásokat **R Markdown** formátumban várjuk, és a tantárgy egy tetszőleges oktatójának kell elküldeni. A feladatmegoldás mind az RStudio szerveren (http://batman.tmit.bme.hu/rstudio), mind saját gépen végezhető.

A feladatokban szereplő adathalmazok vagy nyilvánosan elérhető adathalmazok a megadott URL címeken, vagy a következő, tantárgyi linken érhető el:
https://drive.google.com/drive/folders/1DHywDl4CIwdSPwgb7FBZGclNw2Ue95fV?usp=sharing

Feladatlista

1. **Az ábrázolás nyelvtana**. Az R támogatja a “ggplot2”, vagy “tidyverse” csomagban elérhető ggplot()-ra épülő függvények támogatják a diagrammok (ponthalmazok, idősorok) “ábrázolás nyelvtana” (Grammar of Graphics, GG) szerinti ábrázolását. Mutassa be az GG alapvető felépítését, majd gyakorlatias példákon keresztül mutassa be a ggplot() használatának legáltalánosabb lehetőségeti. (max. 10 pont)
2. **Irreguláris idősorok elemzése.** Mutassa be, hogy az “xts”, vagy “zoo” csomagok miként támogatják az irreguláris idősorok elemzését és predikcióját az RStudio-ban. (max. 15 pont)
3. **Idősorok sztochasztikus becslési módszerei.** Mutasson néhány jellegzetes mószert az idősorok sztochasztikus becslésére. Az egyes módszereket a lényegi elemei, jellegzetességei alapján mutassa be, valamint adjon használatukra példát R-ben. (max. 15 pont)
4. **LDA algoritmus bemutatása és alkalmazása (esszé és gyakorlati feladat).** Mutassa be a Linear Discriminant Analysis (LDA) algoritmus, és vázolja fel, miben különbözik a főkomponens analízis (PCA) módszertől! Az LDA algoritmust mutassa be az cukorbetegség predikciós adathalmazon (https://www.kaggle.com/vikasukani/diabetes-data-set)! Az eredményt értékelje! (12 pont)
5. **Correspondance Analysis (esszé és gyakorlati feladat)**. Mutassa be a Correspondance Analysis módszert, vázolja fel, hogy miben különbözik a főkomponens analízistől (PCA) és mikor célszerű használni. Mutassa be a módszer használatát a COVID-megfázás-influenza tünet adathalmazon (https://www.kaggle.com/walterconway/covid-flu-cold-symptoms)! Az eredményt értékelje és ábrázolja! (15 pont)
6. **ROC görbe (esszé és gyakorlati feladat).** Mutassa be a Receiver Operating Characteristic (ROC) görbét, alkalmazásának célját és módját! Mutassa be a görbe használatát a gyakorlatban az Iris adathalmazon, annak minden egyes jellemzőjére! Melyik jellemző és milyen határérték mellett a legjobb a döntési algoritmus? (8 pont)
7. **Bayes-hálók bemutatása és alkalmazása (esszé és gyakorlati feladat)**. Mutassa be az általános Bayes-hálózatokat, és szerepüket az osztályozási feladatokban! Miben különböznek a naív Bayes-hálózatoktól? A `bnlearn` csomag használatával mutassa be a Bayes-hálózatok használatát az utaselégedettségi adathalmazon (https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction), a hálózat felépítéséhez használja a max-min hill climbing módszert! Számítsa ki, hogy egy 42 éves férfi utas,

akinek a gépe 10 percet késett, mekkora valószínűséggel lesz elégedett az légitársasággal! Tipp: az adathalmazban található folytonos változókból készítsünk kategóriákat megítélésünk szerint! (15 pont)

8. **Pénzügyi adatok elemzése.** Végezze el a “finreportzh.csv” adathalmaz egyes változók és a változók közötti összefüggések elemzését, valamint ábrázolja az egyes iparágakban az egy főre eső profitot boxplot segítségével. Azonosítsa a kiugró értékeket. (max. 10 pont)
9. **Pénzügyi adatok tisztítása.** Végezze el a “finreportzh.csv” adathalmaz tisztítását. Az adattisztítás során minden lépését indokolja. (max. 5 pont)
10. **Ingatlanok adatainak elemzése.** Elemezze, majd keressen összefüggéseket az ingatlan (ingatlan.csv) adathalmazon a lakások különféle paraméterei, valamint a lakások ára között! Ehhez használja fel az adatsorok elemzése, valamint a regressziószámítás elemeit. A feladat eredményeképpen válasszon az egyes paraméterek közül (többet) és határozza meg a kiválasztott paraméterek és a lakások ára közötti többszörös regressziót. Az eredményt értékelje! (Az egyes adatok és a lakások ára között nem feltétlen van értelmezhető összefüggés, ugyanakkor az összefüggés nem biztos, hogy lineáris. Az elemzés során legyen kreatív!). (max. 10 pont)
11. **Regresszió financiális adatokon.** Az egyes iparágakban, a “finreportzh.csv” adathalmazon határozza meg, hogy az egyes adatok között milyen összefüggések léteznek. A vizsgálatokat regressziószámítás módszerével végezze el. Az elért eredményeket értékelje. (max. 5 pont)
12. **Föld melegedésének elemzése.** Végezze el a relatív föld felszíni hőmérséklet (GISTEMP), valamint a hőmérsékleti anomáliák (GCAG) adatait tartalmazó adathalmazon (gisstemp.csv) az idősorok (mind a GISTEMP, mind a GCAG adatok) determinisztikus elemzését. A determinisztikus modellen alapulva mutassa meg, hogy a 2020-as évben milyen értékek a legvalószínűbbek az egyes hónapokban. (max. 5 pont) (Az adathalmaz A legvalószínűbb érték a determinisztikus modell alapján prediktált értékekből adódik!)
13. **Légi utasok.** Végezze el az AirPassangers idősor determinisztikus elemzését. Határozza meg a trend, a szezonális, a véletlen, de még a ciklikus változás komponenseket is. A determinisztikus modell alapján határozza meg, hogy mennyien utaztak repülőgépen 1980-ban. (max. 10 pont)
14. **Elemezze Elon Musk 2017-ig publikált Twitter bejegyzéseit!** (Például: Milyen témákról beszélt gyakran? Gyakori témáit milyen környezetben emlegette? Csoportosíthatók a bejegyzései témák szerint vagy idő szerint?) Az elemzésről készítsen grafikával és szöveges leírással rendelkező elemzést! Az adathalmaz *elonmusk_tweets.csv* néven található. Tipp: a tweet bejegyzések számos hivatkozást (url és nevet) tartalmaznak, ezek szűrésére végezze el! (max. 10 pont)
15. **Elemezze egy olasz kisvárosban kihelyezett levegőminőség mérő eszköz által szolgáltatott adatokat**! Egy olasz kisvárosban két eszközzel, egy tanúsított referencia mérőeszközzel és egy olcsóbb MOS szenzorral (PT08) végeztek levegőminőség méréseket. Milyen összefügéseket vél felfedezni az egyes gázok koncentrációja között? Mely napokon volt a levegő minősége kifogásolható? Milyen minőségű a tesztelt szenzor, a referenciamérésekhez viszonyítva? (max. 10 pont)
    a. Az adathalmazt az *air_quality.csv* fáljban találjuk. Az egyes oszlopok jelzik, hogy referencia mérésről (GT - Ground Truth) vagy a tesztelt szenzortól (PT08) érkező mérésről van szó. *Vegye figyelembe, hogy a hiányzó méréseket a -200 érték jelöli!*
    b. Az adathalmazról részletesebb leírás találhat a https://archive.ics.uci.edu/ml/datasets/Air+Quality# oldalon.
16. **Elemezze az ehető és mérges gombák tulajdonságait, keressen asszociációs szabályokat a gombák tulajdonságai alapján!** A gombák tulajdonságai betűkkel kódolva vannak az adatbázisban (gombak.zip, agaricus-lepiota.data, agaricus-lepiota.names), minden sor egy gomba különböző tulajdonságát jelöli. A mellékelt fájlban megtalálhatók az oszlopok leírásai is. Melyek azok a tulajdonságok amelyek nagy valószínűséggel jelen vannak a mérges gombáknál? Van-e olyan gombacsoport ami nem mérgező? (max. 10 pont)
17. **Elemezze a minőségi borok tulajdonságait!** Előbb keressen outliereket az adatbázisban, majd keressen közös tulajdonságokat a jó borokra. Jó bornak tekinthető például a 7 fölötti quality érték. Az adatbázisban (borok.zip, winequality-red.csv, winequality-white.csv, winequality.names) külön vannak a fehér és vörös borok, illetve mellékelve van a különböző oszlopok elnevezése is. Az elemzést elég elvégezni a fehér borok adatbázisra. (max. 10 pont)
18. **Készítsen osztályozót az 1994-es népszámlálási adatokra**, amely jól meg tudja becsülni, hogy egy adott

tulajdonságokkal rendelkező személy éves fizetése nagyobb, vagy kisebb, mint 50e dollár. Az adatokat az adult.data fájlban találja. Ha szüksége van rá, használhatja a magyar oszlopneveket az adult.names.hu.csv fájlból. (max. 10 pont)
    a. Végezzen előfeldolgozást az adatokon, itt szűrje ki a helytelen értékeket, valamint foglalkozzon a NA értékekkel is. Értékelje az osztályozó teljesítményét különféle tanult mérőszámok szerint. Próbáljon ki többféle osztályozót (legalább 3)!
    b. Bővebb információ az adatokról: http://archive.ics.uci.edu/ml/datasets/Adult
19. Készítsen osztályozót, amely hatékonyan fel tudja ismerni sejtek néhány vizsgált tulajdonsága alapján a rosszindulatú mell daganatot. Az adathalmazt a breast-cancer-wisconsin.data alatt találja. A magyar oszlopneveket a breast-cancer_hu_colnames.csv fájban találja. (max. 10 pont)
    a. Végezzen előfeldolgozást az adatokon, itt szűrje ki a helytelen értékeket, valamint foglalkozzon a NA értékekkel is. Értékelje az osztályozó teljesítményét különféle tanult mérőszámok szerint. Próbáljon ki többféle osztályozót (legalább 3)!
    b. Bővebb információ az adatokról: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
20. Készítsen osztályozót, mely meghatározza egy autó megbízhatóságát különböző tulajdonságai alapján. Az adatokat a car.data fájl tartalmazza. A magyar oszlopneveket a car_hu_colnames.csv fájban találja. (max. 10 pont)
    a. Végezzen előfeldolgozást az adatokon, itt szűrje ki a helytelen értékeket, valamint foglalkozzon a NA értékekkel is. Értékelje az osztályozó teljesítményét különféle tanult mérőszámok szerint. Próbáljon ki többféle osztályozót (legalább 3)!
    b. Bővebb információ az adatokról: http://archive.ics.uci.edu/ml/datasets/Car+Evaluation