Házi feladat választás

A tantárgy teljesítéséhez egy önálló adatelemzési házi feladat kidolgozása és leadása szükséges. A hallgató feladata, hogy egy adathalmaz alapján átfogó adatfeltárást (EDA) és modellalkotást végezzen validációval, majd az eredményeket dokumentálja.

A házi feladat elvégezhető az oktatók által javasolt adathalmazok egyikével, vagy saját választású adathalmazzal. A Teams felületen egy külön feladatban kell megadni a választott témát és adatforrást, egy rövid szöveges dokumentum formájában. Ennek három elemet kell tartalmaznia:

1. Az adatszerkezet rövid leírása vagy linkje (amennyiben általunk kiadott adathalmazt választotok, elegendő a neve és sorszáma),
2. Az adatok és a téma rövid bemutatása (1–2 bekezdés),
3. A vizsgálandó hipotézisek vagy összefüggések felsorolása. A hipotézisek a későbbiekben módosíthatók, de fontos, hogy a kiindulási irány és cél már itt egyértelmű legyen.

Határidő: a téma és adathalmaz kiválasztásának határideje október 22. (szerda) éjfél.

Feltáró adatelemzés (EDA)

A házi feladat első szakaszában áttekintő jellegű adatfeltárást kell készíteni. A hallgató feladata az adatok szerkezetének, eloszlásainak, összefüggéseinek bemutatása, az anomáliák, hiányzó adatok és meglepő tények azonosítása, valamint a további elemzésre érdemes összefüggések kijelölése. Az eredményeket egy Jupyter notebook formájában kell dokumentálni, amely tartalmazza: - a használt kódot, - a kimeneti ábrákat és táblázatokat, - az eredmények szöveges értékelését. Ez a rész az aláírás megszerzéséhez szükséges. A Teams felületen külön leadási pontot biztosítunk, ahova a notebookot és (saját adathalmaz esetén) az adatszerkezetet kell feltölteni.

Határidő: a 12. oktatási hét vége.

Modellalkotás és kiértékelés

Az EDA eredményei alapján ki kell választani legalább egy, modellezésre alkalmas összefüggést vagy hipotézist. A hallgató feladata:

1. az adattisztítás és előkészítés elvégzése,
2. megfelelő modell kiválasztása és betanítása (például regresszió, osztályozás stb.),
3. a modell kiértékelése és az eredmények értelmezése,
4. a hipotézis alátámasztása vagy cáfolása az adatok alapján.

A modellezés során az órákon bemutatott eszközökön túl egyéb módszerek is alkalmazhatók, azonban minden esetben elvárás a megfelelő adatelemzési módszertan következetes betartása. A végső házi feladatot az EDA-részhez hasonlóan Jupyter notebook formátumban kell elkészíteni “report”-ként, amely tartalmazza a felhasznált kódot, a kimeneti ábrákat és táblázatokat, valamint az eredmények szöveges értékelését. A teljes házi feladatot a hallgató vizsgája előtti héten kell leadni. A pontos határidővel és a leadás módjával kapcsolatban a félév során adunk további tájékoztatást.

Az elemzés minden lépése legyen áttekinthető és újrafuttatható. Az alkalmazott módszerek és eszközök megválasztását indokolni kell. A házi feladat önálló munka, de az oktatókkal konzultációs lehetőség

biztosított. Nyelvi modell használata megengedett, azonban itt is a kar irányelvei érvényesek, lásd: https://vik.bme.hu/hallgatoknak/altalanos/mi-hasznalat-ajanlasok

1. Spotify and Youtube

Adat: https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

Ötletek:

- Összefüggés van-e az egyes zenék hangneme (key) és egyéb jellemzők (pl. táncolhatóság, energia, stb.) között? Megbecsülhető-e a hangnem egyéb jellemzők alapján?
- Befolyásolják-e a különböző jellemzők (tempó, ének, stb.) a hallgatottságot? Ha igen, mely jellemzők milyen mértékben?
- Unalmas, hosszú szám alatt többen kommentelnek? Vagy esetleg több like-ot kapott szám alatt?
- (Haladó) A címek alapján tudunk-e tipikus témákat találni? Milyen egyedi - anomáliának minősülő - címeket találunk?

2. IMDB TV shows dataset

Adat: https://www.kaggle.com/datasets/asaniczka/full-tmdb-tv-shows-dataset-2023-150k-shows

Ötletek:

- Fedezze fel a tévéműsorok népszerűségének trendjeit és fő tényezőit.
- Elemezze a tévéműsorok műfajait a legnépszerűbb műfajok vagy műfajkombinációk azonosítása érdekében.
- Vizsgálja meg a tévéműsorok nézettsége és az évadok és epizódok száma közötti kapcsolatot.
- Jósolja meg egy tévéműsor sikerét olyan jellemzők alapján, mint a szavazatszám, az átlag és a népszerűség.
- Azonosítsa a legtermékenyebb tévéműsor-alkotókat vagy produkciós cégeket az általuk készített műsorok száma alapján.
- Fedezze fel a tévéműsorok futási idejének eloszlását, és vizsgálja meg, hogy az epizódok időtartama befolyásolja-e az általános nézettséget.
- Vizsgálja meg a tévéműsorok gyártásának trendjeit a különböző országokban és hálózatokban.
- Elemezze a tévéműsorok nyelve és népszerűsége közötti kapcsolatot, és vizsgálja meg a nem angol nyelvű műsorok népszerűségét.

3. Global Salary Data

Adat: https://www.kaggle.com/datasets/zedataweaver/global-salary-data

Ötletek:

- Érdemes megvizsgálni fizetési trendeket, egyenlőtlenségeket és eltéréseket az egyes országok, iparágak és munkakörök között. Azonosítani a legmagasabb átlagfizetéssel rendelkező vagy a versenyképes fizetési ajánlatokkal rendelkező országokat.
- Végezzen statisztikai elemzéseket, hogy betekintést nyerjen a fizetéseloszlásokba, a standard eltérésekbe és a kiugró értékekbe. Vizsgálja meg a jövedelmi egyenlőtlenségeket és azok kapcsolatát a regionális gazdasági tényezőkkel.

4. 10,000 Data about movies

Adat: https://www.kaggle.com/datasets/willianoliveiragibin/10000-data-about-movies-1915-2023

Ötletek:

- Érzelemelemzés a felhasználói értékelések és vélemények alapján.
- Milyen összefüggések vannak az egyes változók között (értékelés-műfaj, játékidő-műfaj, bevétel-műfaj, bevétel-besorolás, stb.)
- Trendelemzés különböző évtizedekben. Hogyan változtak a filmek tulajdonságai az évek során (műfaj, játékidő, besorolás)

5. Labdarúgó játékosok átigazolások adatbázisa – market value

Adat: transfermarkt_market_value.csv

Adat eredete: https://www.transfermarkt.com/

Az egyik legnépszerűbb nyilvánosan elérhető platform, amely labdarúgó játékosok átigazolási adatait tartja nyilván a Transfermarkt, ahol a látogatók megtalálhatják a kulcsfontosságú jellemzőket és a tényleges átigazolási díjakat vagy a becsült piaca értékeket az egyes játékosok esetében. A Transfermarkt a játékosok értékének becslésére egy modellt vezetett be, amely a ,,közösség bölcsességére" (wisdom of the crowd) hagyatkozik.

Ötletek:

- Transfermarkt modelljének értékelése, tehát a market value és valós átigazolási díj vizsgálata
- (Haladó) Gépi tanulási/mély tanulási modell építése, mely jobban vagy megközelítőleg, olyan jól teljesít, mint a Transfermarkt modellje.

6. Labdarúgó játékosok átigazolások adatbázisa

Adat: transfermarkt_atigazolasok.csv

Adat eredete: https://www.transfermarkt.com/

Ötletek:

- Átigazolási trendek vizsgálata (ország, bajnokság, életkor, stb.)
- Mely jellemzők befolyásolják legjobban az átigazolási díjat?

7. Kerékpár bérlés

Adat: https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset

Ötletek:

- Függ a bérlések száma az időjárástól? Mire a legérzékenyebb (hőmérséklet, eső, stb.)?
- Az év adott időszakától függ a regisztrált és eseti felhasználók száma?
- A regisztrált felhasználók munkába járáshoz használják a kerékpárokat? Az eseti felhasználók szórakozni? (pl. milyen idősávokban)
- Becsüljük meg, hogy adott időjárás esetén mennyien fogják a kerékpárokat kibérelni!

8. EU statisztikák - szabad feladat (haladó)

Adat: https://ec.europa.eu/eurostat/data/database

Az EUStat weboldaláról számtalan adatsor letölthető, amelyek alkalmasak lehetnek érdekes hipotézisek ellenőrzésére. Például:

- Az emberek szubjektív egészségi állapotának érzete összefügg-e az ország egészségügyi rendszerének állapotával és költségvetésével?
- Igaz, hogy alacsonyabb GDP-vel rendelkező országokban nagyobb mértékben növekedtek az ingatlanárak?

A feladat célja az adatok áttekintése, majd saját hipotézisek megfogalmazása. A feladat nehézsége, hogy aggregált adatok nem tölthetők le, tehát saját adathalmazt kell építeni az elérhető adatsorokból.

9. Ütközési adatok

Adat: https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes

Ötletek:

- Tipikusan hány autó ütközik egy balesetben? Milyen az eloszlása az ütköző autók számának?
- Melyek a leggyakoribb okok az ütközésekben?
- Vajon az időjárás befolyásolja a balesetek számát (haladó)?
- Miként lehetne csoportosítani a baleseteket? Vannak tipikus csoportok?

10. Ingatlan árak

Adat: ingatlan.csv (Teams csoport feltöltései között)

Ötletek:

- Miként alakultak az ingatlan árak adott kerületekben?
- Helyileg mekkora eltérések vannak az ingatlanok árai között (haladó: csoportosítsuk területenként)?
- Van összefüggés az eladási ár és az egyes jellemzők között?

11. Tőzsdei elemzés

Adatok: stocks_data

A fentebbi adathalmaz a három legnagyobb amerikai tőzsdeindexbe (S&P500, Dow Jones Industrial Average, Nasdaq-100) cégeinek historikus részvényárfolyamait tartalmazza. Emellett maguknak az indexeknek a teljesítménye is megtalálható.

Ötletek:

- Nominális kontra reálhozam: cégek teljesítményének kompenzációja az infláció mértékével (pl. normalizálás 2000-es USD-ra)
- Nyerő portfólió: állítson össze olyan részvényportfóliót, amellyel az indexek teljesítményénél nagyobb hozamot lehetett elérni, 10, 20, 30, 40, 50 éves időszakra vonatkozóan!
- Béta-együttható kiszámolása egyes részvényekre (piaccal, indexszel való korreláció)

- Együttmozgó részvények: megfigyelhető, hogy bizonyos szempontból egy csoportba tartozó (azonos szektor, hasonló termékek gyártása) részvények árfolyamai között (akár pozitív, akár negatív) korreláció figyelhető meg. Keressen ilyen részvénypárokat, -csoportokat!
    - Anomáliák keresése: felfedezhető-e korreláció olyan részvények között is, melyeknél első látásra nem lenne indokolt? Van-e ezekre mégis valami magyarázat?
- Szektorok teljesítménye:
    - Hogy teljesítenek az egyes szektorok az indexekhez képest különböző időszakokban?
    - Szektorrotáció vizsgálata: Az elmélet szerint a különböző szektorok különböző időszakban teljesítenek jól. Az adatok alapján igazolható ez az elmélet? Van magyarázat arra, hogy melyik szektor mikor teljesített jól vagy rosszabbul?
- Az adathalmazban található részvények összevetése egyéb piacokkal: Európai/Ázsiai tőzsdék, kötvénypiacok teljesítménye, stb.

12. Labdarúgó játékos metrikák adatbázis (FBREF)

Adat: https://docs.google.com/spreadsheets/d/1lQgIDcxsHT1m_IayMldmiHVOt4ICbX-ys8Mh9rggPHM/edit#gid=2093389914

Adat eredete (példa): https://fbref.com/en/players/6ce43701/Youssoufa-Moukoko

Ötletek:

- Jellemzők közötti összefüggések feltárása (pl: xG –Goal, xA-Assist, stb.)
- Játékosok klaszterezése: A szakdolgozat célja a labdarúgók szerepének és profiljának jobb megértése. A játékosok kategorizációja a pozíciójuk alapján nem a legjobb módszer. A hallgató feladata játékosok klaszterezése a csapatukon belüli szerepének értékelésére, túllépve a merev pozíciós besorolásokon.