Ki nyeri a következő Diáknapokat?

Alkalmazott gépi tanulás projekt
Mesterséges Intelligencia tantárgyból

🔖 Bevezetés:

Ebben a feladatban egy alkalmazott gépi tanulás projektet fogsz megvalósítani, amely a csapatok teljesítményének összehasonlítására fókuszál a kolozsvári KMDSZ DIáknapokon. Az Diáknapok során Kolozsvár a “világ közepe”, ahol a különböző baráti csoportoktól verbuvált csapatok szórakoznak együtt, de ugyanakkor versenyeznek is egymással több, mint 50 próbán. Történelmi adatok elemzésével egy gépi tanulás modellt fogsz készíteni, hogy elemezd a csapatok teljesítményét, felderítsd azokat a tényezőket, amelyek hozzájárulnak sikerükhöz - és talán megjósold a következő nyertest.

A legjobb beadandókat - beleegyezés esetén - megosztjuk a KMDSZ irodával, illetve publikáljuk a Székelydata adatvizualizációs blogon.

📚 Feladat lebontása:

📊 1. Adatgyűjtés és előfeldolgozás / Data wrangling:

a. Összegyűjtöttük azokat az adatforrásokat, amelyek releváns információkat tartalmaznak a csapatok teljesítményéről az Diáknapokon. Ide tartozhatnak olyan adatrészletek, amelyek információkat tartalmaznak az egyes csapatok által szerzett pontok számáról különböző Diáknapokon. **Négy mappát találsz** a projekthez: *pontszamok*, *programok*, *nyers html* és *logok*

b. A *pontszamok* mappa tartalmazza az egyes csapatok eredményeit, évente gyűjtve. Bizonyos években megvannak a részletes lebontások az egyes próbákra, más években csak összpontszám, vagy csak a csapatlista áll rendelkezésre. Az adatokat nem normalizáltuk, nyersen kapod.

c. A *programok* mappa tartalmazza a részletes programok menetrendjét, pontszámait. Ez csak azon évek esetében releváns, ahol nincs meg a teljes próbánkénti lebontás. Az adatokat nem normalizáltuk, nyersen kapod - itt nagyon sok utómunkára van szükség, hogy használható legyen, rád van bízva, hogy eldöntsd hoz-e elegendő hozzáadott értéket a projekthez.

d. A *nyers html* az Internet Arhívumról elmentett html oldalt tartalmazza.Ebből lettek kinyerve a fenti adatok, így nem kulcsfontosságú - de nyugodtan böngészd, ha hasznosnak találod.

e. A *logok* mappa egyes kiadások grafikai elemeit tartalmazza

💫 **2. Normalizálás / Data normalization:**

a. Az adatokat előfeldolgozd a hiányzó értékek kezelésével, csapatnevek, próbanevek és a numerikus értékek normalizálásával, kategorikus változók kódolásával és szükség esetén adattisztítási műveletek elvégzésével.

b. Itt arra kell gondolj, hogy hasznos lehet egy bizonyos csapat részvételét és régiségét nyomon követni, még akkor is ha nincsen meg a pontszámlebontás egy adott évben. Ugyanez igaz a próbákra is, meg talán a napszakra, helyszínekre is.

🌍 **2. Felfedező adatelemzés / Data discovery:**

a. Végezz felfedező adatelemzést az adatokkal kapcsolatos betekintések megszerzéséhez. Elemezd a trendeket, mintákat és az pontszámok eloszlását az egyes csapatok és Diáknapos kiadások között.

b. Vizualizáld az adatokat megfelelő diagramokkal, például oszlopdiagramokkal, vonaldiagramokkal, eloszlásokkal, hogy kiemeld az érdekes megfigyeléseket és megkönnyítsd az csapatok közötti összehasonlításokat.

🌟 **3. Jellemzők kialakítása / Feature engineering:**

a. Hozz létre új jellemzőket (neveztük dimenziók/oszlopoknak is) az elérhető adatok alapján, amelyek növelhetik a modell prediktív erejét. Ide tartozhatnak lineárisan vagy one-hot módszerrel kódolt változók, mesterségesen létrehozott időbélyegek.

b. Szabadon behozhatsz más léíró adatokat is - például ha ismered bizonyos csapatokat alkotó baráti csoportok forrásvárosait, adhatsz szubjektív “cool-factor”-t, mérhetsz médiavisszhangot, like-számot, stb. Ez a lépés nem szükséges a feladat elvégzéshez és teljes mértékkel opcionális.

🏗️ **4. Modellépítés / Data modeling:**

a. Osztd fel az előfeldolgozott adatokat tanító- és teszthalmazra. Két feladatot oldjál meg:

b. Válassz egy felügyelt tanulási / supervised learning algoritmust az csapatok teljesítményének előrejelzésére, például lineáris regresszió - linear regression, döntési fák / decision tree, véletlen erdők / random forest. Az sklearn python könyvtárból bármelyik algoritmust használhatod.Használhatsz osztályozási módszereket / classification is a rangsor helyett - például valaki bekerül-e a top 5-be vagy sem.

c. Válasz egy felügyeletlen tanulási algoritmust / unsupervised learning és próbáld megvizsgálni, hogy van-e szabályszerű rendeződés a csapatok között / clustering. Hogy segítsen az eredmények elemzésében, használhatsz egy dimenziócsökkentési algoritmust / dimenionality reduction hogy ki tud plottolni 2D-ben a kialakított adat-klasztereket.

c. Taníts be összesen minimum 2 különböző modellt a tanítóhalmazon (legalább egy felügyelt és egy felügyeletlen) és hasonlítsd össze a különböző modellek teljesítményét, és azonosítsd a legjobb eredményt nyújtót. Ez egy alsó határ, kipróbálhatsz több algoritmust.

💬 **5. Modell értelmezés és következtetések / Conclusions:**

a. Értelmezd a tanított modellt annak megértéséhez, hogy mely tényezők járulnak hozzá az csapatok sikeréhez az Diáknapokon.

c. Kutass érdekes összefüggések után a jellemzők (nagyrészt próbák) és az csapatok sikerének között, és mutasd be ezeket világos és tömör módon.

📜 **6. Beadandó:**

a. A feladatot bármilyen programozási nyelvben elkészítheted. Mi eddig *python*-t használtunk, de használhatsz *R*-t, *Matlab*-ot vagy bármi mást.

b. Használhasz bármilyen gépi tanulási könyvtárat. Mi eddig *sklearn*-t használtunk, de használhatsz *keras*-t, *TensorFlow*-t, *PyTorch*-ot vagy akár nagy nyelvi modelleket is, mint a *ChatGPT* vagy a *Huggingface* modelljei.

c. Készíts részletes *Jupyter* vagy *Google Cloab* munkafüzetet, amely dokumentálja a felfedezéseidet, a módszertanodat és az eredményeket. Ez működő kód kell legyen és hiba nélkül lefusson egy teljesen új környezetben is.

d. Ez egyetlen fileban tartalmazza a vizualizációkat, táblázatokat és releváns statisztikákat az elemzésed alátámasztásához. Használj kód és szöveg cellákat felváltva a magyarázatod elősegítésére.

e. Foglald össze röviden (300-500 szó) a következtetéseidet a csapatok teljesítményével kapcsolatban az Diáknapokon, az elkészült gépi tanulás modell előrejelzéseire és következtetéseire alapozva. Azokra a kérdésekre próbálj választ adni, hogy “Ki nyeri a meg 2024-es diáknapokat” illetve mely csapatok viselkednek “klaszter”-ként. Ugyanakkor a szervező KMDSZ csapat felé is próbálj megfogalmazni javaslatokat - mely próbák hasznosak, döntik el a végeredményt.

f. A moodle-re egyetlen file-t vagy linket tölts fel, ami a .ipynb file-t tartalmazza, vagy az elő munkafüzetre mutat. **FONTOS: ha Google Colab-ot használsz, ne felejtsd el a munkafüzetet PUBLIKUS-ra állítani!**

g. Ne felejtsd el legalább minimálisan dokumentálni a kódodat: a cél az, hogy egy hozzád hasonló technikai színvonalon levő felhasználó megértse az eredményeidet és az elemzést.

🎖️ Értékelési szempontok:

1. Adatgyűjtés és előfeldolgozás: 25%
2. Felfedező adatelemzés: 25%
3. Jellemzők kialakítása: 10%
4. Modellépítés és értékelés: 10%
5. Modell értelmezése és következtetések: 10%
6. Beadandó formája és minősége: 20%

Megjegyzés:
A fenti lebontás általános útmutatóként szolgál. Szabadon módosíthatod vagy bővítheted a feladatot az egyéni megértésed és szakértelmed alapján.

Osztályozás:
Tudatában vagyok, hogy ez egy **nehéz** projekt. De úgy gondolom, hogy egy valós adatokat tartalmazó projekt a legjobb módja e tantárgy alkalmazásának - a maga kihívásaival együtt. Ezzel együtt, az erőfeszítést és a logikus gondolatmenet és adatelemzést részesítem előnyben és nem a nyers eredményeket. Ezt a megközelítést a pontszám-lebontás is tükrözi.

És ne feledd:

```
from sklearn.modell_csalad import ModellNeve
model = ModellNeve()

// supervised
model.fit(X, y)
y_pred = model.predict(X)

//unsupervised
model.fit(X)
model.transform(X)
```

Sok sikert a projekthez! 🚀🚀🚀

Dénes