Nagy mennyiségű adatfeldolgozás (INBGM9935E)

(projektmunka)

1. Minden hallgató a saját adathalmazát feldolgozva önállóan készíti el a projektmunkát.
2. Az elemzés során tetszőleges programozási nyelv és eszköztár használható (a laborgyakorlatokon megismert Jupyter Noteboook fejlesztői környezet és Pandas, Seaborn, Matplotlib, Scikit-learn könyvtárak használata előnyös lehet).
3. A projektmunka során az alábbi A, B, C és D részekben megfogalmazott feladatokat kell megoldani.
4. Minden olyan feladat esetén, ahol attribútumok kiválasztására van szükség, indokolja meg a döntést, illetve a kiválasztás célját.

A.) Adatvizualizáció és klaszterezés

A1.) Az adathalmaz megismerése, a benne szereplő attribútumok bemutatása és általános jellemzése
A2.) Előfeldolgozás, adattisztítás, pl. hiányzó adatok vagy extrém értékek feltérképezése, illetve kezelése, adatkonverzió
A3.) Különböző vizualizációs eszközök használata az adathalmaz, illetve az attribútumok között fennálló kapcsolatok feltárására, két/több attribútum együttes vizsgálata alkalmas plotok felhasználásával és a kapott eredmények értelmezése
A4.) Két különböző, klaszterezésre használható algoritmus kiválasztása, rövid bemutatása, illetve azoknak az adathalmazra történő alkalmazása, a kapott eredmények értelmezése (szükség esetén pl. a Scikit-learn dokumentáció is használható).
Az A3.) feladat esetén legalább öt különböző vizualizációs eszközt használjon fel! A kapott eredmények közül emelje ki azokat, amelyeket a további feladatok során (pl. az attribútumok kiválasztásánál) később is fel tud majd használni!

B.) Lineáris regresszió

B1.) Az előző rész eredményeit felhasználva, a megfelelő attribútumok kiválasztása és azok indoklása, a vizsgálandó regressziós feladat megfogalmazása

B2.) Lineáris regresszió alkalmazása folytonos attribútum esetén, az eredmények értelmezése, illetve azok felhasználása

C.) Logisztikus regresszió

C1.) Az előző rész eredményeit felhasználva, a megfelelő attribútumok kiválasztása és azok indoklása, a vizsgálandó osztályozási feladat megfogalmazása

C2.) Logisztikus regresszió alkalmazása diszkrét attribútum esetén, az eredmények értelmezése, illetve azok felhasználása

A B2.) és C2.) feladatok esetén legalább két különböző modellt vizsgáljon meg!

D.) További osztályozási módszerek, visszatartó gépi tanulás

D1.) Az előző részek eredményeit felhasználva, a megfelelő attribútumok kiválasztása és azok indoklása, a vizsgálandó osztályozási feladat megfogalmazása

D2.) Előfeldolgozás, pl. a kiválasztásra került attribútumok értékeinek normalizálása

D3.) A tanuló algoritmus használatához az adathalmaz tanuló és teszthalmazra történő felbontása, a választott paraméterértékek indoklása

D4.) Két, a logisztikus regressziótól különböző, osztályozásra alkalmas algoritmus kiválasztása, azok rövid bemutatása, az adathalmazon elvégzett tanítás eredménye, a kapott paraméterértékek bemutatása, értelmezése

D5.) Kiértékelés, különböző teljesítménymértékek használata