

ADATELŐKÉSZÍTÉS

(3) Data Preparation

Adatok tisztítása (Data Cleaning)

Transzformáció

Redukció

(A) ADATOK TISZTÍTÁSA

- A szükségtelen és hibás adatok eltávolítása az adatok közül
- Redunanciák
 - Azonos adatok, vagy objektumok kiszűrése és törlése
- Adatintegritás
 - Ellentmondások, inkonzisztenciák feltárása és feloldása
- Hiányzó értékek
 - ...
- Hibás értékek
 - Formailag hibás értékek kezelése
 - Outlierek
 - El kell döntni, hogy az outlier-ek fontos részét képezik-e a modellalkotásnak
 - Anomáliák

(B) TRANSZFORMÁCIÓ

- A rendelkezésre álló adatok transzformálása a hatékonyabb modellalkotás érdekében
- Új adatok bevezetése
 - A modellalkotáshoz szükséges új adatok létrehozása
- Adatok normalizálása
 - Adatok egységes értékkészletre hozása
- Folytonos értékek diszkretizálása
 - Szükséges értékkészlet előállítása

(C) REDUKCIÓ

- A feldolgozandó adatmennyiség csökkentése a tárolási, számítási komplexitás, csökkentésének a kezelhetőség növekedésének érdekében
- **Attribútumok számának csökkentése**
 - Tulajdonságok elhagyása
 - Tulajdonságok összefésülése → Dimenziócsökkentés
- **Adatmennyiség csökkentése**
 - Mintavételezés
 - Adatok csoportosítása

HIÁNYZÓ ÉRTÉKEK KEZELÉSI STRATÉGIÁK

- Egyértelmű adatok megadása
 - pl. megye településhez
- Adatok becslése
 - Átlaggal
 - Mediánnal
 - a medián nem érzékeny kiugró értékekre
 - Korrelációk és hasonlóságok alapján
 - Modell alapján töltjük ki → Regresszió – Predikció
- Adat törlése
 - Az adathalmaz kevésbé lesz szignifikáns
 - pl. nem kritikus adat esetében és más megoldás nem jöhet szóba
- Dummy érték bevezetése
- Változtatás nélkül
 - pl. nem fontos az analízis szempontjából