Rövid beszámoló

Elsősorban egy pilot alkalmazást kellett megcsinálnunk, minden részt önállóan, hogy betekintést nyerjünk minden részlegbe. A kapott feladat viszonylag egyszerű volt, Hana adatbázisból kapott adatokat vizualizálni tetszőleges front-end technológiával, illetve pár extra funkció implementálása, mint például az adatok szűrése.

Bár a feladat egyszerűnek hangzik, hozzá rengeteg mindent kellett megismerni. Először is magával a Hana környezettel, hogyan működik, hogyan lehet használni. Ezután következett a tábla létrehozása, adatokkal feltöltése, illetve utána normalizálása és az adatok tisztítása, majd az egészet egy kalkulációs nézetbe helyezése. Illetve utána Angular-ral vizualizálás ODATA segítségével. Ezek által kis belátást nyertem modellezési, illetve front-end témakörökbe.

Ezután történt a szakosodás, amiben az adatbányászatot választottam Hana környezetben, azon belül a PAL eljárások. Mivel teljesen új témakör volt ez számomra, rengeteg segédanyagot kellett elolvasnom magáról a témáról. Általában egy-egy algoritmus családról bevezetőket, melyik mire használható, milyen eredményeket ad, azok értelmezése, illetve milyen következményeket lehet levonni azokból. Továbbá maguk az algoritmusok egy családon belül, részletesebben a működésükről, melyiket mikor érdemes, miért érdemesebb mint a többit használni az adott esetben.

Négy adatbányász témakörrel foglalkoztam a legtöbbet: klaszterezés (pl. K-Közép, AP, DBSCAN), osztályozás (pl. K Legközelebbi Szomszéd), regressziószámítás (Lineáris Regressziók), illetve asszociációk (pl. FP-Növés, Apriori). Ezeket tanulmányoztam Hana környezetben, hogyan lehet használni őket. Illetve rengeteget kipróbáltam több adathalmazra is, különböző paraméterekkel, figyelve az eredmény változását.

Legutolsó sorban az eddig gyűjtött tudást felhasználva végeztünk egy összehasonlító elemzést az adatbányászatról két külön környezetben: PAL és Python. Természetesen közös adathalmazon történt az összehasonlítás. Először is kezdtük az algoritmusok testreszabhatóságával, melyik környezetben milyen paraméterekkel lehet módosítani az eljárás működését. Utána a futási időt mértük össze, illetve az eredményeket. Rengeteg féleképpen futtattuk az algoritmusokat, különböző paraméter kombinációkkal amik segítségével látványosabb különbségeket tapasztaltunk.

Majd az összehasonlításból származó tapasztalatokat rendszereztük, összefoglaltuk és levontuk a következtetéseket belőle. Elsősorban egyes környezetek előnyei, illetve hátrányairól szól, kitérve az adott feladat elvégzéséhez szükséges követelményeire, illetve használhatóságára.