Rövid beszámoló

Kiss Alex

1 Pilot projekt

A szakmai gyakorlat első hete alatt megismerkedtünk a **SAP HANA** környezettel a hivatalos dokumentációk és útmutatók segítségével. A környezet megismeréséhez az útmutatókat követve létrehoztam egy **HTML5** modult, valamint egy **Java** projektet is. Ezt követően létre kellett hoznunk egy pilot-projekt keretében egy, az **Országleltár** weboldalához hasonló oldalt, aminek back-end oldali megoldásaként a HANA környezet szolgált. Ehhez kaptunk egy *.csv* állományt, melyet az adatbázisba való feltöltés előtt normalizálnunk kellett. Ezt először a HANA környezetben próbáltam megoldani, de mivel rengeteg duplikátum, illetve hibás adat szerepelt az állományban, inkább egy script segítségével hoztam a megfelelő alakra. Ezt követően a webes felületen definiáltam a megfelelő sémát, majd azt az adatoknak megfelelően finomhangoltam.
Ahhoz, hogy a front-end megoldás el tudja érni az adatokat, létre kellett hoznunk egy **OData** modult. Ehhez a hivatalos útmutatót követtem, melynek keretében jobban megismerkedtem a REST API-k használatával. A front-end megoldásként az **Angular** rendszert választottuk, melynek megismeréséhez a hivatalos útmutatót, valamint a **W3Schools** segédanyagait olvastam. A front-end részeként a weboldalon elhelyeztem három táblázatot, melyeket az OData modulon keresztül feltöltöttem a régiókkal, a megyékkel és a településekkel, a felhasználó által kiválasztott szűkítés szerint. Ezeket az adatokat egyes attribútumok szerint csoportosítva a **CanvasJS** könyvtár egy grafikonján jelenítettem meg.

2 Adatbányászat

A pilot projektet követően egy megbeszélés keretében megállapodtunk, hogy ki melyik modul tagjaként szeretné folytatni a szakmai gyakorlatot. Elsődleges modulként az adatbányászatot választottam **Python** környezetben, másodlagos modulként pedig a front-end fejlesztést **Angular** környezetben.

Mivel még korábban nem használtam Python-t, így az első napokban különböző útmutatók segítségével megismerkedtem a nyelv szintaktikájával, a vezérlési szerkezetekkel, az alapvető adatszerkezetekkel, valamint a modulok használatával. Ezt követően el kezdtem az adatbányászat témakörével foglalkozni. A témakör elméleti részét főleg a Bevezetés az adatbányászat című könyvből, a gyakorlati részét pedig a Learning Data Mining with Python című könyvből ismertem meg.

Az elkövetkezendő hetek során különböző felügyelt (K-legközelebbi szomszédok módszer, Logisztikus regresszió, Naív-Bayes módszer, Szupport vektor gép) és felügyelet nélküli osztályozási eljárásokat (K-közép módszer, DBSCAN), asszociációs, regressziós és klaszterező algoritmusokat vizsgáltam meg, ehhez saját, illetve a **SciKit** modul implementációit használva. Az algoritmusok implementálásához, illetve a már meglévő modulok implementációinak mélyebb megértéséhez egyes modulokkal is jobban megismerkedtem, mint például a **NumPy** illetve a **Pandas**, továbbá az adatok megjelenítéséhez a **matplotlib**-bel. Az adatvizualizáció megkönnyítése érdekében a **Jupyter Notebook**-ot használtam.

A szakmai gyakorlat utolsó hetében az egyik **PAL**-t használó csapattársammal választható feladatként egy összehasonlító elemzés elkészítésének keretében összevetettük a két környezetet. Ehhez ki kellett választanunk egy már *"bevált"* adathalmazt, ám a HANA környezetben az importáláskor fellépő problémák miatt egy kisebb állományt kellett választanunk, így az írisz-adathalmazra esett a választás. Egyes technikai problémák miatt a kívántnál kevesebb algoritmust tudtunk összehasonlítani, így csak a **DBSCAN-t**, a **K-közép módszert** valamint a **K-legközelebbi szomszédok módszert** vetettük össze. Az algoritmusoknak megvizsgáltuk a paraméterezhetőségeit, eredményeit és futásidejüket, továbbá az összehasonlítás részeként a két környezet lehetőségeit az adatok importálására, valamint az eredmények megjelenítésére.