Mesterséges intelligencia beadandó

**Adatok betöltésének kihívásai**

Az adatok betöltése során többféle probléma is felmerült, amelyek megoldása a projekt korai szakaszában elengedhetetlen volt. Az adathalmaz különböző típusú oszlopokat tartalmazott, amelyek között nem numerikus és hiányos értékek is szerepeltek. A nem numerikus oszlopokat először szűrnünk kellett, majd a megfelelő kódolási technikákat alkalmazva (pl. LabelEncoder) átalakítottuk őket olyan formátumra, amelyet a modellek kezelni tudnak. Ez a lépés elengedhetetlen volt a modellek pontosságának növelése érdekében.

A duplikált sorok problémája szintén jelentős nehézséget jelentett. Ezek a redundanciák torzíthatták volna az elemzési eredményeket, ezért ezeket a sorokat eltávolítottuk az adathalmazból. Az adatok tisztítása során külön figyelmet fordítottunk a hiányos sorokra is: ezeket vagy elhagytuk, vagy statisztikai módszerekkel pótoltuk (pl. medián vagy átlag alapján).

**Adathalmaz méretének kezelése**

A projekt során felhasznált adathalmazunk mérete rendkívül nagy volt: a 126 ezer soros tanító adathalmaz jelentős memóriaigényt támasztott. Ez nemcsak a betöltésnél, hanem az elemzések és modellezés során is problémát okozott. Például a gépi tanulási modellek betanítása a nagy adathalmaz miatt időigényes volt, különösen összetettebb modellek esetében. A probléma mérséklése érdekében csökkentettük a modellezési paramétereket, például a döntési fák maximális mélységét és a használt jellemzők számát.

Az időigényes betanítási folyamatokat monitoroztuk a Python time könyvtárával, amely lehetővé tette az egyes lépések futási idejének pontos mérését. Ezek az adatok segítettek optimalizálni a modell paramétereit, hogy az elérhető erőforrások mellett a legjobb teljesítményt nyújtsa.

**Modell teljesítményének javítása**

A modellek pontosságának növelése érdekében különböző technikákat alkalmaztunk. A Recursive Feature Elimination (RFE) módszer segített kiválasztani a legfontosabb jellemzőket, amelyek tényleges hatással voltak a modell teljesítményére. Ez nemcsak csökkentette a modellezés bonyolultságát, hanem hozzájárult a számítási idő csökkentéséhez is. A keresztvalidáció segítségével pedig biztosítottuk, hogy a modellek stabilan és megbízhatóan teljesítsenek az ismeretlen adatokon is.

A különböző metrikák, mint például a pontosság, az F1-score és a ROC-görbe, lehetőséget adtak a modellek alapos kiértékelésére. Ezek az eredmények nemcsak az aktuális modellezés sikerességét mérték, hanem iránymutatást adtak a további fejlesztésekhez is.

**Anomáliák azonosítása**

Az anomáliák felderítése különösen összetett feladat volt, amely során távolsági alapú és klaszterezési módszereket használtunk. Az optimális klaszterszám meghatározása során az "Elbow" módszert alkalmaztuk, amely vizuálisan segített megtalálni a legjobb beállítást. Az anomáliákat a centroidoktól való távolság alapján azonosítottuk, ahol a küszöbértéket az átlagos távolság plusz kétszeres szórás határozta meg. Ez a megközelítés hatékonyan kiszűrte a szokatlan értékeket, amelyek torzíthatták volna a modellezési eredményeket.

**Adatok vizualizálása**

A nagy adathalmaz vizualizálása kihívást jelentett a komplexitás és a változatosság miatt. Az oszlopok közötti eltérések és a nagy mennyiségű adat kezelése megnehezítette a mintázatok felismerését. Ennek érdekében különféle vizualizációs eszközöket (pl. heatmap, barplot, countplot) használtunk, amelyek segítettek az adatok közötti kapcsolatok és anomáliák ábrázolásában. Ezek az ábrák nemcsak a modellezés előkészítésében voltak hasznosak, hanem az eredmények kommunikációját is megkönnyítették.

**Webes alkalmazás fejlesztése**

A Streamlit segítségével készített webes alkalmazásunk célja az volt, hogy az eredmények könnyen hozzáférhetők és átláthatók legyenek. Az alkalmazás felépítése során nagy hangsúlyt fektettünk az intuitív kialakításra: az adatok vizualizációját és a statisztikai mutatókat külön szekciókba rendeztük, hogy a felhasználók egyszerűen navigálhassanak közöttük. Fontos tapasztalat volt, hogy a Streamlit alkalmazás megfelelő működéséhez szükség van a scripts mappa aktiválására, hogy a fejlesztői környezetből zökkenőmentesen elérhető legyen.