JELEN PUBLIKÁCIÓ A NEMZETI ADATGAZDASÁGI TUDÁSKÖZPONT KÖZREMŰKÖDÉSÉVEL JELENT MEG.

Adatelemzési folyamat és keretrendszer a közigazgatás számára

BOGACSOVICS GERGŐ, HAJDU ANDRÁS, HARANGI BALÁZS, LAKATOS ISTVÁN, LAKATOS RÓBERT, SZABÓ MARIANNA, TIBA ATTILA, TÓTH JÁNOS, TARCSI ÁDÁM *

Absztrakt: A mesterséges intelligencia utóbbi évtizedben bekövetkezett ugrásszerű fejlődése az azt támogató hardveres és szoftveres platformok folyamatos bővülésével az adatelemzést is új szintre emelte. Ez a szintlépés alapvetően úgy értelmezhető leginkább, hogy egyre kevésbé szükséges a feldolgozó modellek precíz meghatározása, mivel már a most rendelke-

* **Bogacsovics Gergő:** PhD tanulmányait 2020-ban kezdte meg a Debreceni Egyetem Informatikai tudományok Doktori iskolájában. Kutatásai területei közé tartoznak: mesterséges intelligencia, gépi tanulás, mélytanulás, mély megerősítéses tanulás, orvosi képfeldolgozás és természetes szövegfeldolgozás.

Prof. Dr. Hajdu András: 2003-ban szerezte Ph.D., 2017-ben MTA doktori fokozatát. IEEE senior member, jelenleg az Informatikai Kar dékánja, az Informatikai Doktori Iskola Adattudomány és vizualizáció programjának és a Képfeldolgozás és Komputergrafika Tanszék vezetője. Elsődleges kutatási területe a diszkrét matematika, digitális képfeldolgozás, gépi tanulás, nagy mennyiségű adatfeldolgozás.

Dr. Harangi Balázs: 2015-ben szerezte meg Ph.D. fokozatát, melyhez kapcsolódó disszertációjának témája orvosi képfeldolgozás volt. Elsődleges kutatási témái közé tartoznak olyan területek, mint az orvosi képfeldolgozás, textúra elemzés, gépi tanuló rendszerek és mesterséges intelligencia (mélytanulás).

Lakatos István: 2020-től Ph.D. hallgatóként végzi kutatását a Debreceni Egyetem Informatikai Tudományok Doktori Iskolájában az Adattudomány és vizualizáció doktori programban. Elsődleges kutatási témája a gyógyszertervezés és molekula szintézis elősegítése gépi tanuló rendszerek és neurális hálózatok által.

Lakatos Róbert: 2020-tól a Debreceni Egyetem Informatika Karán, mint Ph. D. hallgató folytat tanulmányokat és végez kutatómunkát a mesterséges intelligencia, gépi tanlás és a szöveg feldolgozás területén. Továbbá tanulmányai mellet több mint 10 éve folyamatosan dolgozik szoftver fejlesztőként.

Szabó Marianna: 2019-től Ph.D. hallgatóként végzi kutatási és oktatói munkáját a Debreceni Egyetem Informatikai Tudományok Doktori Iskolájában, kutatási témája a valószínűségi időjárás előrejelzések statisztikai utófeldolgozása, de érdeklődései közé tartozik a neurális hálózatok és az adatelemzés tágabb témaköre is.

Tiba Attila: 2020-tól tanársegédként dolgozik az Informatikai Kar Komputergrafika és Képfeldolgozás Tanszékén. Fő érdeklődési területei: sztochasztikus optimalizálás, gépi tanulás, digitális képfeldolgozás, numerikus matematika, nagy mennyiségű adatfeldolgozás.

Tóth János: Jelenleg a Debreceni Egyetem Informatikai Tudományok Doktori Iskolájának doktorjelöltje. Elsődleges kutatási területei közzé tartozik az ensemble rendszerek optimalizációja, az orvosi képfeldolgozás és a gépi tanuló rendszerek.

Tarcsi Ádám: Az Eötvös Loránd Tudományegyetem Informatikai Karán, az Adattudomány és Adattechnológiai tanszéken felel az ipari kapcsolatokért, valamint a Nemzeti Adatgazdasági Kutatóközpont szakmai divízióvezetője. Kutatási területei közé tartozik: vállalati információs rendszerek, adattudományok.

zésre álló eszközök képesek biztosítani, hogy pusztán a nyers input adatok megfelelő szolgáltatásával és az elérni kívánt cél meghatározásával az effektív elemzést végző eljárás – általában neurális háló architektúra – már egy gépi tanulási folyamaton keresztül automatikusan kerüljön kialakításra. Mivel ez a trend a jövőben várhatóan tovább fog erősödni, az elemzési eljárásokat célszerű úgy felépíteni, hogy ebbe a keretrendszerbe illeszkedjenek. Ennek megfelelően hangsúlyt kell fektetni a feldolgozni kívánt, potenciálisan különféle területekről származó adatbázisok olyan előfeldolgozására, amelyet követően a teljes adatkészlet átadható az elemző architektúrának. Mivel az elemzés eredményének értelmezhetőségét emberi felhasználásra is alkalmassá kell tenni, ezért tipikusan vizualizációs technikákat alkalmazhatunk erre a célra. Értelemszerűen a vizualizációs technikát is a hatékonyság miatt a teljes elemzési keretrendszerbe érdemes integrálni, azaz a vizualizációs eszköz közvetlenül ráépül az elemzőarchitektúra kimenetére, illetve annak belső adatábrázolására, amennyiben például a bemeneti adatok közötti összefüggések bemutatása is hasznos a döntéshozás indoklásához.

Kulcsszavak: adatelemzés, keretrendszer, adatelemzési terv

1. BEVEZETÉS

Az alábbi táblázat bemutatja a sikeres adatelemzés megvalósításához szükséges lépéseket a letisztult CRISP-DM[1,2,3] és ASUM-DM[4] elemzési módszertanokból kiindulva. Az effektív adatelemzés megkezdésekor kiemelt figyelmet érdemel az adatok előkészítése, majd a megfelelő elemzőmodell kiválasztása. Az elemzési feladat összetettségétől függően az egyes feladatok kidolgozásának mélysége értelemszerűen igen széles skálán mozog. Ezek a szintek alapvetően az adatelemzés következő négy leggyakoribb feladattípusának megfelelően azonosíthatóak be: leíró jellegű elemzés/riport vagy megfigyelés; feltáró elemzés; mély elemzés prediktív analitikával kiegészítve; egyedi elemzés, tanulmány. Számos esetben az elemzési feladat alapvetően egy jelentéssel és értelmezést segítő fejlesztéssel zárul. Komolyabb fejlesztést is igénylő megrendelés esetén a végtermék elkészítésénél a dőlttel szedett pontok is relevánssá válnak.

Célok meghatározása/ Üzleti megértés	**Megrendelői célok megértése** • Háttér • Üzleti folyamatok • Üzleti elvárások **Helyzetfelmérés** • Az erőforrások felmérése • Követelmények, feltételezések, korlátozások • Kockázatok és váratlan események • Terminológia • Költségek és előnyök **Adatelemzési feladat megfogalmazása** • Adatelemzési célok • Adatelemzés sikerének kritériumai **Az adatelemzési terv elkészítése** • Projektterv • Eszközök és technikák előzetes felmérése
Adatok megértése	**Adatok összegyűjtése** • Kiinduló adatgyűjtési jelentés **Adatok leírása** • Adatleíró jelentés **Adatok feltárása** • Adatfeltáró jelentés **Az adatminőség ellenőrzése** • Adatminőségi jelentés
Adatok előkészítése	**Adatok kiválasztása** • A felvétel indoklása/Kizárás **Adat minőség javítás** • Adat minőség jelentés **Adatszármaztatás** • Származtatott attribútumok • Generált rekordok **Adatok integrálása** • Egyesített adatok **Adatok formátumának módosítása** • Átformázott adatok • Adatkészlet • Adatkészlet leírása

Modellalkotás az elemzéshez	**Modellezési technikák kiválasztása** • Modellezési technika • Modellezési feltételezések **Tesztelési módszertan kialakítása** • Teszttervezés **Modell építése** • Paraméter-beállítások • Modellek • Modell leírások **Modell kiértékelése** • Modell értékelése • Paraméter-beállítások finomítása
Kiértékelés	**Eredmények kiértékelése** • Adatelemzés eredményének kiértékelése üzleti siker kritériumokra nézve • Jóváhagyott modellek **Módszertan, felülvizsgálati folyamat** • A folyamat felülvizsgálata **Következő lépések meghatározása** • A lehetséges intézkedések listája • Döntés
Végtermék elkészítése	**Eredmények megjelenítése** • Megjelenítő felület tervezése **Zárójelentés készítése** • Zárójelentés • Végleges bemutató **Végtermék tervezése** • Telepítési és bevezetési terv **Monitoring és karbantartás tervezése** • Monitoring és karbantartási terv **A projekt felülvizsgálata** • Tapasztalati dokumentáció

2. CÉLOK MEGHATÁROZÁSA

Ez a kezdeti szakasz a projekt céljainak és követelményeinek megértésére összpontosít egy adatelemzési problémadefiníció és egy adatelemzési terv kidolgozásának érdekében. Ennek keretében elkészül egy költségbecslés – beleértve a kidolgozás idejét is – a feladat összetettségének függvényében.

Az adatelemzés célja, hogy a megrendelői igény a lehető legjobban legyen teljesítve. Ehhez szükséges a megrendelő részéről felvetett probléma megértése, valamint az, hogy annak megoldására reális célok kerüljenek kitűzésre.

Az igénykielégítés során konzultációk keretében kerülnek a részletek pontosításra. A konzultáció egy iteratív folyamat, így az elemzési cél is folyamatosan tisztul. Ennek során a megrendelő számára egyértelművé válik, hogy mit várhat el az elkészülő adatelemzéstől. A konzultációs folyamat végére meghatározásra kerülnek az igénylendő adatkörök is.

2.1. MEGRENDELŐI CÉLOK MEGÉRTÉSE (A KONZULTÁCIÓ RÉSZEKÉNT)

E lépés célja annak megértése, hogy valójában mire szeretne a megrendelő választ kapni, milyen eredményt szeretne kézhez kapni, milyen formátumban és mennyi időn belül. Azaz itt szükséges az elvárt eredmény pontos meghatározása, az elvárások specifikálása, valamint a felhasználni kívánt adatok körének előzetes meghatározása is. E lépés során fel kell tárni azokat a tényezőket, amelyek meghatározzák az adatelemzési projekt jellegét, magas szintű, általános kérdésekkel.

- *Az eredménytermék:* Milyen formátumban szükséges elkészíteni az elemzést? Szöveges tanulmány, szöveges elemzés, részletes prezentáció, áttekintő prezentáció, adatvizualizáció módja, interaktív felületekre való igény, estelegesen szolgáltatás és hozzátartozó felhasználói felület?
- *Az elemzési feladatra rendelkezésre álló idő:* Mikorra szükséges elkészíteni az elemzést?
- *Gyakoriság szempontok:* Egyszeri elemzés készül, többszöri, esetleg folyamatos elemzési szolgáltatás, folyamatos információ igény áll fenn megrendelő oldaláról? Megismételhető-e az elemzés – adatok elérhetősége vagy más korlátok miatt egyáltalán lehet-e szó az elemzés megismétléséről? Igényt tarthat-e a megrendelő erre és milyen gyakran?

2.2. HELYZETFELMÉRÉS (A KONZULTÁCIÓ RÉSZEKÉNT)

A helyzetfelmérés célja annak meghatározása, hogy mi valósítható meg az adatelemzéssel. Ehhez részletesebb tényfeltárás szükséges a feldolgozandó adatok köréről, az azokkal kapcsolatos esetleges megkötésekről és feltételezésekről, amelyeket figyelembe kell venni az adatelemzési terv meghatározásakor.

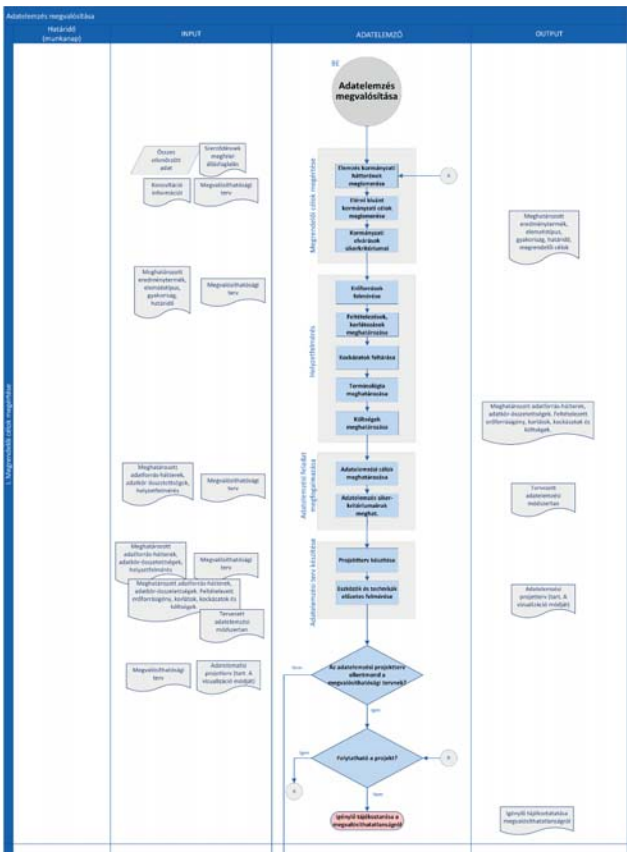
- *Adatforrások, azok háttere, összetettsége:* Az elemzéshez szükséges adatok és azok forrásainak beazonosítása.
- *Felhasználandó adatkörök és az adatkészletek összetettsége:* Milyenek az elemzéshez szükséges adatkörök, mekkora az adatkészletek és a becsatornázandó adatforrások, illetve az érintett szervezetek száma?
- *Elemzés típusa:*
  - Leíró jellegű elemzés/riport vagy megfigyelés: Az érintett adatkészleteknek általános jellemzését hivatott megadni, amihez alapvetően a leíró statisztika eszközrendszerét használjuk.
  - Feltáró elemzés: A szükséges adatkészletek vizsgálata hagyományos statisztikai, elemzési eszközökkel, amelyek alkalmasak az adatokban rejlő összefüggések feltárására, okok vizsgálatára, összehasonlítására.
  - Mély elemzés prediktív analitikával kiegészítve: Az elemzés átfogó képet ad az igényelt nagy mennyiségű és komplexebb adatkörök tekintetében. A multidiszciplináris elemzésünk részeként a tématerületet vizsgáljuk jogi, közigazgatási, közgazdasági, valamint műszaki szempontok alapján. Az elemzés részét képezi prediktív analitikai előrejelzés, amely segíti a jövőbeli döntéshozatalt a megismert információk alapján. A mesterséges intelligenciával történő elemzéseket, predikciókat ágazati szakértők bevonásával végezzük el.
  - Egyedi elemzés, tanulmány: A beérkezett igények szerint a fenti három kategóriába egyértelműen nem besorolható, vagy több elemzési módszert is felvonultató, mély szakmai ismeretet igénylő és számos területet, adatkört érintő elemzés, nagyobb részt egyedi formátumban, méretben, elemzési mélységben és tartalommal készíthető el.

2.3. ADATELEMZÉSI FELADAT MEGFOGALMAZÁSA

E lépésben a megrendelői célok kerülnek átfogalmazásra adatelemzési projekt célkitűzésekké, valamint meghatározásra kerülnek az adatelemzés tervezett eredménytermékei, illetve az elemzés sikerességének technikai kritériumai.

2.4. AZ ADATELEMZÉSI TERV ELKÉSZÍTÉSE

A folyamat e szakaszában az adatelemzési célok – és ezáltal a megrendelői célok – eléréséhez szükséges terv elkészítése történik meg. A terv tartalmazza az adatelemzési projekt további szakaszaiban végrehajtandó lépéseket, beleértve az adatok összegyűjtését és az alkalmazandó eszközök és módszertanok kiválasztását is.



1. ábra: Célok meghatározása

3. ADATOK MEGÉRTÉSE

Az adatok megértése az igényelt adatok összegyűjtésével indul, és olyan tevékenységekkel folytatódik, amelyek lehetővé teszik az adatok megismerését és az adatminőségi problémák azonosítását. Ennek során vizsgálatra kerül az összegyűjtött adatok formátuma, mérete, belső struktúrája, valamint megtörténik a minőség ellenőrzése (azaz a hiányzó értékek, hibák és az inkonzisztencia kiszűrése) az adatokban.

3.1. ADATOK ÖSSZEGYŰJTÉSE

E lépésben a korábban azonosított adatcsomagok összegyűjtése és a további elemzéshez szükséges rendszerezése, adatbázisba rendezése történik meg. Ennek a lépésnek azokban az esetekben van relevanciája, amikor a kulcsszolgáltató által előkészített adatcsomagok struktúrája nem megfelelő a további feldolgozás szempontjából. Megjegyzendő, hogy ez a lépés nagymértékben túlmutathat az adatok egyszerű beszerzésén, mivel az átadott adat igen változó formátumban lehet (strukturálatlan dokumentum, űrlap-jellegű PDF állományok, képek, szövegek stb.). Ezért itt további (ETL – Extract – Transform – Load – azaz adat betöltés az adatforrásból, adattranszformáció, adatbetöltés) fejlesztésekre lehet szükség.

3.2. ADATOK LEÍRÁSA

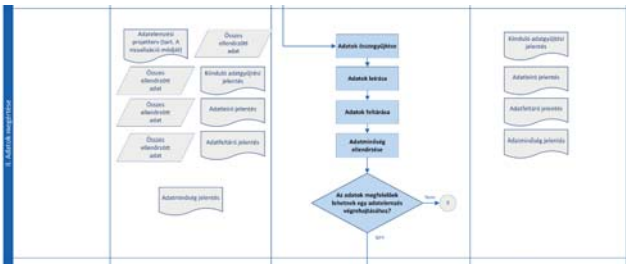
Ebben a lépésben az összegyűjtött adatok „felszíni" tulajdonságainak leírása történik meg, azaz megvizsgálásra kerül az adatok formátuma, mennyisége (például a rekordok/mezők/adatelemek száma), kódolási sémája stb., valamint kiértékelésre kerül, hogy az adatok megfelelnek-e a követelményeknek.

3.3. ADATOK FELTÁRÁSA

Ez a szakasz adatelemzési szempontból tárja fel a rendelkezésre álló adatokat különböző vizualizációs és statisztikai módszerek segítségével. Ezek közé tartoznak a kulcsattribútumok (például egy osztályozási feladat esetében az osztálycímkék) eloszlása, az attribútumok közötti kapcsolatok feltárása, az egyszerű aggregációk eredményei, a jelentős részpopulációk tulajdonságai stb. Ezek az elemzések hozzájárulnak az adat minőségének vizsgálatához, pontosítják az adatok leírását, illetve segítenek a további adatelőkészítési lépések végrehajtásában is.

3.4. AZ ADATMINŐSÉG ELLENŐRZÉSE

Ebben a lépésben az adatok minősége kerül vizsgálatra a következő szempontok alapján:
Teljesek-e az adatok (minden szükséges esetre kiterjednek-e)? Helyesek-e az adatok? Ha tartalmaznak hibákat, azok milyen jellegűek és mennyire gyakoriak? Vannak-e hiányzó értékek az adatokban? Ha igen, hogyan vannak reprezentálva, hol fordulnak elő, és mennyire gyakoriak?



2. ábra: Adatok megértése

4. ADATOK ELŐKÉSZÍTÉSE

Az adatok előkészítése magában foglal minden olyan tevékenységet, amely az elemzés során használt adatkészletek létrehozásához szükséges a kezdeti nyers adatokból. Az adatelőkészítési feladatokat esetleg többször is el kell végezni az alkalmazott modellek függvényében.

Ebben a szakaszban megtörténik a feladat megoldásához szükséges adatcsomagok előfeldolgozása, beleértve a nagy mennyiségű adatok kezelésére alkalmas adatbányászati eszközökkel történő rendszerezést, adatbázisba szervezést is. Ezen tevékenység során figyelmet kell fordítani az adatok előfeldolgozására, a megfelelő formátum kialakítására, az adatok validálására és egységesítésére is.

4.1. ADATOK KIVÁLASZTÁSA

Ebben a lépésben kerül eldöntésre, hogy a rendelkezésre álló adatok mely része kerül majd felhasználásra az elemzés során. A kritériumok közé tartozik az adatelemzési célok relevanciája, az adat minősége és az esetleges technikai korlátok, például az adatmennyiség vagy az adattípusok korlátai.

4.2. ADATMINŐSÉG-JAVÍTÁS

A folyamat e szakaszában az adatok minőségének a kiválasztott elemzési technikák által megkövetelt szintre javítása történik meg. A javítás alapjául az előző szakasz adatminőség-ellenőrzési lépése során talált minőségi problémák szolgálnak. Az adattisztítás magában foglalhatja az adatok tiszta részhalmazainak kiválasztását, az adatokra vonatkozó séma definiálását, illetve annak megléte esetén a megfelelő validálását és alapértelmezések beillesztését, a duplikátumok eltávolítását, az adattípusok átalakítását, a gépelési hibák javítását stb., vagy akár a hiányzó adatok modellezéssel történő becslését is, ha szükséges.

4.3. ADATSZÁRMAZTATÁS

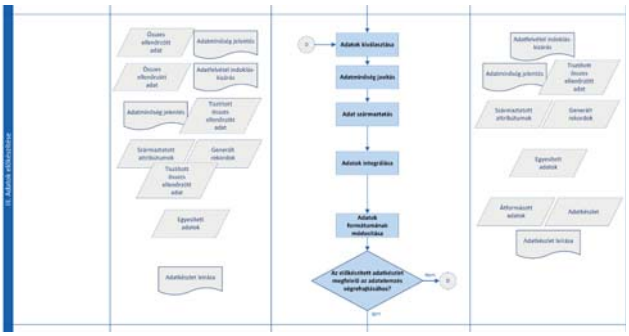
Ebben a lépésben olyan konstruktív adatelőkészítési műveletek kerülnek végrehajtásra, amelyek a tárolt attribútumokból származtatott attribútumokat vagy teljesen új adatelemeket állítanak elő. Ide tartozik a meglévő attribútumok értékeinek transzformálása is.

4.4. ADATOK INTEGRÁLÁSA

Az adatok integrálása során a különböző adatforrásokból származó adatok kombinálása történik meg. Erre akkor van szükség, ha több olyan adatforrás van, amely vagy ugyanazokról az objektumokról tartalmaz különböző információkat (összekapcsolás), vagy hasonló információkat tartalmaz, de különböző objektumokról (hozzáfűzés).

4.5. ADATOK FORMÁTUMÁNAK MÓDOSÍTÁSA

A formátummódosítás során olyan transzformációk kerülnek végrehajtásra az adaton, amelyek nem változtatják meg az adat jelentését, de a modellező eszköz számára szükségesek lehetnek, például az attribútumok sorrendjének megváltoztatása, az attribútumok címkéinek módosítása.



3. ábra: Adatok előkészítése

5. MODELLALKOTÁS AZ ELEMZÉSHEZ

Ebben a szakaszban különböző modellezési technikákat kell kiválasztani és azokat alkalmazni, valamint a paramétereiket az optimális értékekre állítani. Jellemzően ugyanarra az adatelemzési problématípus megoldására több technika is létezik. Egyes technikáknak lehetnek specifikus követelményei az előkészített adatok formájára vonatkozóan, ezért szükséges lehet visszatérni az adatelőkészítési szakaszhoz.

5. 1. MODELL KIVÁLASZTÁSA

A modellezés első lépéseként ki kell választanunk az elemzéshez illeszkedő modellezési technikát, amelyet használni kívánunk. Jó eséllyel ennek alapvető iránya meghatározásra került a korábbi konzultációs fázisban, viszont most az elemzés típusához megfelelő konkrét eszközöket kell kiválasztani.

Leíró jellegű elemzés/riport vagy megfigyelés esetén:
- Klaszterező megoldások
- Egyszerű statisztikai számítások

Feltáró elemzés esetén:
- Korreláció kutatás
- Ok-okozati kutatás

Mély elemzés prediktív analitikával kiegészítve:
- Osztályozó, illetve regressziós algoritmusok (döntési fa építésére, neurális hálózat generálására)

Több technika alkalmazása esetén ezt a feladatot minden egyes technikára külön-külön kell elvégezni.

5.2. TESZTELÉSI MÓDSZERTAN KIALAKÍTÁSA

Mielőtt ténylegesen megépítenénk egy modellt, ki kell alakítanunk egy módszertant az elemzés eredményének tesztelésére. Például a felügyelt adatelemzési feladatokban (osztályozási feladat) rendelkezésre állnak a gyakori hibaarányokat mérő mutatók, de vannak olyan esetek, amikor a jóság mérésének a módja nehezen determinálható. Milyen módon kívánjuk az elemzés pontosságát mérni? Van-e ismert és elismert mérési módszer? Ki kell-e dolgozni az eredményesség mérésének módszerét? Kell-e külső szakértő a kiértékeléshez? Van-e rendelkezésre álló teszt adat?

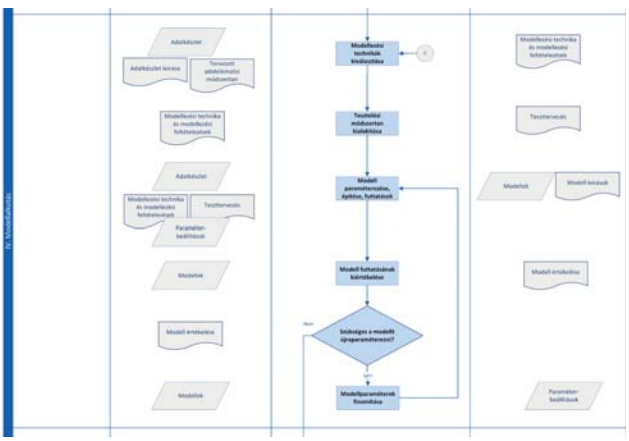
5.3. MODELL ÉPÍTÉSE A RENDELKEZÉSRE ÁLLÓ ADATHALMAZ SEGÍTSÉGÉVEL

Bármely elemzés esetén a választott modellező eszköznek számos paramétere lehet és azokat optimalizálni kell. Meg kell tudni határozni a szükséges és opcionális paramétereket, valamint a paraméterbeállítások kiválasztásának indoklását. Kérdés a paraméterek számossága, amely összefüggésben van az alkalmazott modell komplexitásával. A komplex modellek esetén kérdés a rendelkezésre álló erőforrások alkalmassága, illetve amennyiben szükséges, a felhőben elérhető kapacitás költségbecslése.

5.4. MODELL KIÉRTÉKELÉSE

Az elkészült modelleket a szakterület ismeretei alapján ki kell értékelni, továbbá szükséges felmérni a modellezési és feltárási technikák sikerességét. Itt szükség lehet az üzleti elemzőkre és

a szakterületi szakértőkre, hogy megvitassák az eredményeket az adott kontextusban. A modell kiértékelése során csak magát a modellt vesszük figyelembe, minden más eredményt, amely a projekt során keletkezett, majd az értékelési fázisban tekintünk át. Amennyiben több modell is készült (legtöbbször igen), megpróbáljuk rangsorolni a modelleket. A modelleket az előre meghatározott szempontok szerint értékeljük, így különösen figyelembe kell venni az elkészült modellek számát, illetve azok eredményei alapján az esetleges visszatérés szükségességét egy korábbi fázishoz.



4. ábra: Modellalkotás az elemzéshez

6. A VÉGREHAJTOTT ADATELEMZÉSI MÓDSZERTAN KIÉRTÉKELÉSE

Ebben a szakaszban a korábbi fázisokban felépített és optimalizált modell minőségét határozzuk meg. Mielőtt a modell végleges termékként átadásra kerül a megrendelő számára, fontos, hogy alaposan kiértékeljük és felülvizsgáljuk, hogy a létrehozás során megtett lépéseket valóban rendben hajtottuk-e végre, illetve, hogy a modell megfelelően megvalósítja-e az eredetileg kitűzött célokat. Ennek a lépésnek a végén döntést kell hozni az adatelemzés eredményeinek felhasználásáról és átadhatóságáról.

6.1. A KAPOTT EREDMÉNYEK KIÉRTÉKELÉSE

A korábbi értékelési lépések olyan tényezőkkel foglalkoztak, mint a modell pontossága és általánosító képessége. Ebben a szakaszban azt vizsgáljuk, hogy több kidolgozott modell esetén, melyik milyen mértékben felel meg a konzultáció során megfogalmazott céloknak, és feladat feltárni a korábban megfogalmazott célok alapján az esetleges hiányosságokat. A kapott eredmények kiértékelése során szintén hangsúlyos a megrendelővel történő kapcsolattartás. Fontos döntési pont, hogy több megoldás esetén a magasabb pontosság, a gyorsabb átfutási idő vagy az olcsóbb üzemeltetés számít-e jobban? Vizsgálandó továbbá, hogy létrejött-e egyáltalán az a megoldás, amit a megrendelő eredetileg elgondolt? Itt lehet javaslatokat tenni a jövőbeli kapcsolódó adatelemzési szükségletekre.

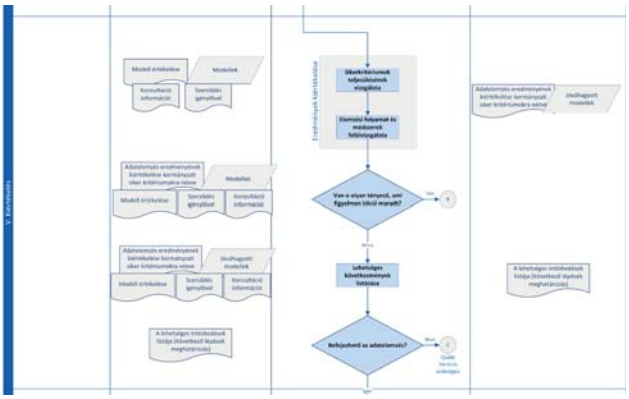
6.2. AZ ALKALMAZOTT MÓDSZERTAN ÁTTEKINTÉSE

Amennyiben eljutunk eddig a pontig, akkor a már elkészült adatelemzési megoldások kielégítőek, és megfelnek a konzultáció során megfogalmazott igényeknek. Utolsó ellenőrzésként szükséges az alkalmazott módszereket felülvizsgálni annak érdekében, hogy kiderüljön, hogy van-e még olyan fontos tényező vagy feladat, amely valahogyan figyelmen kívül maradt. Helyesen épül-e fel a modell? Csak azokat az attribútumokat használja-e, amelyek megengedettek, és amelyek a későbbi elemzésekhez is rendelkezésre állnak majd?

6.3. KÖVETKEZŐ LÉPÉSEK MEGHATÁROZÁSA

A megelőző értékelések és a felülvizsgálat eredményeitől függően döntés születik arról, hogy befejezhető-e az adatelemzéshez kapcsolódó fejlesz-

tés vagy esetleg kezdeményez bármely fél további iterációkat, vagy esetleg új adatelemzési projektet. Ez a feladat magában foglalja a fennmaradó erőforrások elemzését és a költségvetést, amelyek befolyásolhatják a döntéseket.



5. ábra: A végrehajtott adatelemzési módszertan kiértékelése

7. KÉSZTERMÉK ELKÉSZÍTÉSE ÉS ÁTADÁSA

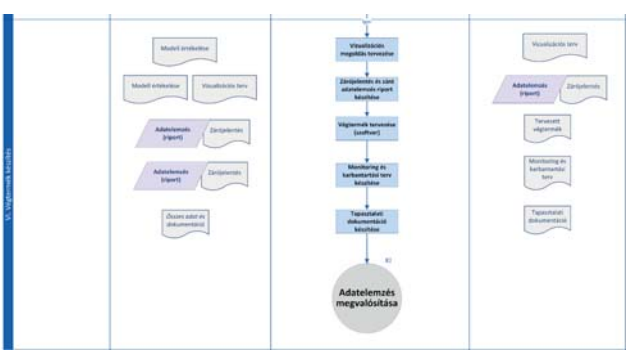
Az adatelemzést megvalósító módszertan önmagában nem lehet egy adatelemzési feladat végterméke. Ahhoz, hogy az végtermékként felhasználható legyen egy esetleges döntéshozásban, a megszerzett és kinyert információkat úgy kell elrendezni és bemutatni, hogy a megrendelő minél egyszerűbben és áttekinthető módon azokat hasznosítani tudja. A követelményektől függően a végső eredménytermék jelentheti egy egyszerű jelentés létrehozását, vagy olyan alkalmazást, amely megismételhető adatelemzések megvalósítására is alkalmas.

7.1. ADATELEMZÉSBŐL SZÁRMAZÓ EREDMÉNYEK MEGJELENÍTÉSE

Az adatelemzés során feldolgozott adatokat és a belőlük kinyert információkat az esetek többségében számszerűsített formában kapjuk vissza. Ezeknek az adatoknak, információknak az értelmezése ebben a formában nehéz és kényelmetlen. Az értelmezés megkönnyítésére egy megfelelő vizualizációs megoldás ad lehetőséget. Ezen a ponton több kérdés is felmerülhet. A vizualizált eredmény lehet egy egyszerű dashboard vagy egy interaktív felület, amelyen akár ki- és bekapcsolhatók az egyes adatszintek. A megoldások között vannak statikus és dinamikus elemek, amelyek az időbeliséggel képesek több információt még értelmezhető módon átadni. Kérdés lehet még, hogy a vizualizáció során alkalmazott eszköz egy meglévő, standard megoldás vagy esetlegesen újat kell fejleszteni.

7.2. ZÁRÓJELENTÉS KÉSZÍTÉSE

A projekt végén a legtöbb esetben az adatelemzés részleteiről és az elért eredményekről zárójelentés készül. Az elemzés mélységétől függően a jelentés lehet csupán a feladat és a tapasztalatok összefoglalása egy rövidebb, néhány oldalas dokumentum, amely a könnyebb értelmezés érdekében diagrammokat, valamint grafikus elemeket tartalmaz. Összetett elemzés esetén érdemes lehet összefoglalni a fejlesztés során szerzett fontos tapasztalatokat (buktatók, félrevezető megközelítések, legmegfelelőbb adatelemzési technikák kiválasztására vonatkozó javaslatok), amelyek szintén részét képezhetik a dokumentációnak, amely egy vezetői összefoglaló formájában kerül átadásra.

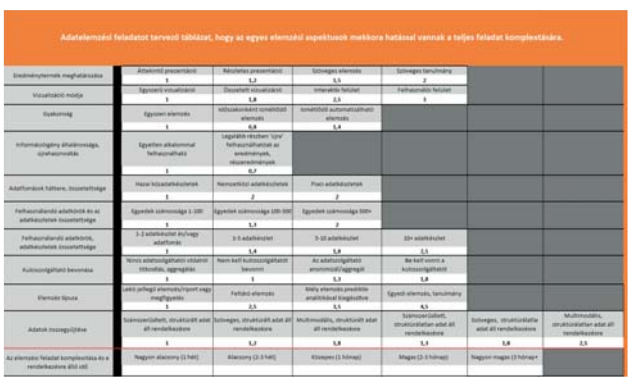


6. ábra: Késztermék elkészítése és átadása

8. AZ ADATELEMZÉS KOMPLEXITÁSÁNAK MEGHATÁROZÁSA

A konzultáció során feltárt aspektusok együttesen határozzák meg az elemzés komplexitását, melynek pontos szintje csak az elemzési munka lezárultával válik ismerté a szereplők számára, ugyanakkor már a konzultációs folyamat elejétől törekedni kell arra, hogy az elemzési komplexitás szintje becsülhető legyen. A becsült komplexitási szint hivatott meghatározni azt, hogy az eljárásrend egyes részeiben kik és milyen tevékenységben vesznek részt, továbbá egy indikációt is jelent a szükséges erőforrások és időszükséglet tervezéséhez.

A feladat komplexitásának becsléséhez segítséget nyújthat az alábbi táblázat kitöltése a konzultáció során feltett és megválaszolt kérdések alapján:



A komplexitás szintje alapján a feladatok az alábbi kategóriákba sorolhatók (becsült időigénnyel):

- Nagyon alacsony (~1 hét időigény)
- Alacsony (~2-3 hét időigény)
- Közepes (~1 hónap időigény)
- Magas (~2-3 hónap időigény)
- Nagyon magas (~3 hónap+ időigény)

8.1. PÉLDA 1: EGYSZERŰ LEÍRÓ ELEMZÉS

Az alábbi elemzés célja az volt, hogy az elérhető adatkörök felhasználásával meghatározzuk a 2020 nyarán megfigyelhető balatoni turizmus főbb célpontjait, trendjeit, területi eloszlását. A covid lezárások után kiemelt cél volt, hogy a lakosság a nyarat belföldön töltse – ennek a célkitűzésnek a csaknem valós idejű „monitoringját” valósítja meg az elemzés. Általában a turizmus statisztikák leginkább a vendégéjszakák adataira épülnek, így csak a szálló vendégek kerülnek bele a mutatókba, ezzel az elemzéssel azonban kimutathatóvá váltak azok a turisztikai célú mozgások is, amelyeket nem szállóvendégek generáltak vagy csak települések között napközbeni „vándorlással” valósultak meg.

Az elemzés során a mozgási adatokat biztosító mobil telefonos lokációadatok, valamint az elemzést és megjelenítést segítendő térképes adatbázisok – OSM térképek – kerültek felhasználásra.

Az elkészült elemzés egy egyszerű dashboard-on került bemutatásra.

A leíró jellegű elemzés célja egy könnyen áttekinthető, ugyanakkor kattintható térképes felület volt, a könnyebb érthetőség és a részletes (település szintű) információk feltárása érdekében. Adatforrások tekintetében lényegében egyetlen adatforrásra támaszkodtunk, ami egy piacról beszerezhető mobil készülékek mozgási adatait tartalmazó adatkészlet volt. Ezt egészítette ki a vizu-

alizációt segítő térképes adatbázis, a települések poligon adataival.

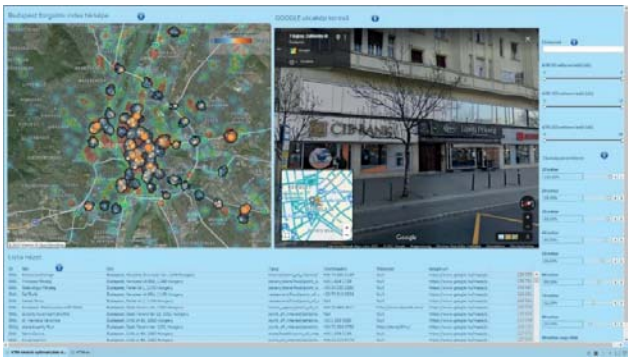
Az elemzés komplexitása a fenti komplexitási mátrix alkalmazásával:

8.2. PÉLDA 2: FELTÁRÓ ELEMZÉS

A megbízás egy olyan elemzés elkészítésére irányult, mely vizsgálja az egyes banki ATM-ek teljesítményét – a napi tranzakciók számát – és a lokációk adottságainak mutatóit:

- a helyben lakó lakosság számosságát és szocio-demográfiai ismérveit,
- a tömegközlekedési információkat,
- a gyalogos és autós forgalom leíró adatait,
- az intézményi ellátottságot és az intézményi látogatottságot.

Az elemzés során lényegében azt vizsgáltuk, hogy az egyes jól tipizálható területeken – belváros, külváros, lakóövezet stb. – melyek azok a meghatározó tényezők, amelyek leginkább korrelációt mutatnak az ATM-ek forgalmi mutatóival.
A feladat során az elemzés eredményeit felhasználva elkészült egy térképes „kereső” felület is, mely alkalmas az egyes kulcs attribútumok súlyozására és így egy interaktív térkép segítségével a felhasználó maga tudja megkeresni a leginkább megfelelő lokációkat egy-egy területen.



Az elemzés komplexitása

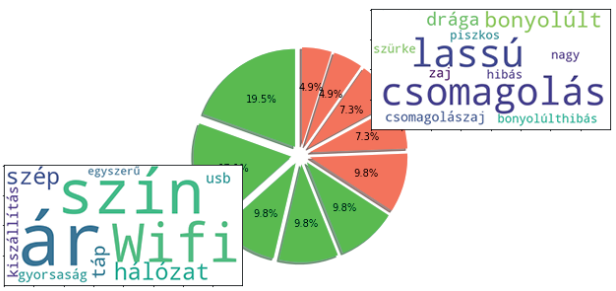
A feladat egy egyszeri feltáró elemzésből állt, mely során meghatároztuk a releváns mutatókat, és ezt követően került kialakításra egy térképes kereső felület, ahol a releváns mutatók súlyozásával lehetett meghatározni a legjobb területeket. Az elemzés során felhasználásra kerültek a fent felsorolt adatkörök, valamint a pénzintézet által rendelkezésre bocsátott ATM tranzakciós adatok. Minden adat „egyszerű” táblázatos formában állt rendelkezésre, így azok előfeldolgozására nem volt szükség. A feltáró elemzési feladat egyszeri volt, egy teljes év adatait felhasználva készült el az elemzés. A feltáró elemzésen alapuló térképes eszköz fejlesztése szintén egyszeri volt, ugyanakkor a mögöttes adatok frissítése folyamatos. A feladat feltáró jellegű elemzési részének eredményei egyszerű prezentáció során kerültek bemutatásra és átadásra a megrendelőnek, hogy a megértés minél gyorsabb és az információk átadása minél érthetőbb módon történjen, míg a térképes kereső már egy interaktív vizualizáció formájában került megvalósításra.

8.3. PÉLDA 3: MÉLY ELEMZÉS PREDIKTÍV ANALITIKÁVAL KIEGÉSZÍTVE

Kis- és középvállalkozások (KKV-k) ellenállóképességének jövőbeli predikciója:
A megfogalmazott projekt célok alapján egy olyan platform kerül kialakításra, ahol a cégek saját vagy más cégek könyvelési (mérleg, eredmény kimutatás) adataikat tudják fel-, illetve betölteni. Továbbá cél egy olyan, a KKV számára elérhető prediktív szolgáltatás, amely képes az adott KKV mérlegek adatainak elemzéséből egy jóslást adni a fizetésképtelenné válás esélyeire vonatkozóan. A mérlegek feldolgozását követően dashboard-on jelenik meg az elmúlt időszak átfogó leírását mutató diagram csoport, amely egyszerűsíti az adatok áttekintését és megértését. Jenleg az utolsó 3 év céges mérlegét elemzi a rendszer. A kiértékeléshez egy olyan gépi tanuláson alapuló modell került kidolgozásra, amely a korábban felszámolásra került céges adatok alapján lett betanítva. A rendszer ennek az MI alapú modellnek a segítségével hívja fel a figyelmet a fizetésképtelenség veszélyét jelző anomáliákra. A fejlesztés során az adatforrás PDF állományokban eltárolt mérlegadatok kinyerésén alapult, majd azok felhasználásával különböző közgazdasági mutatók kerültek származtatásra. Ezt követően gépi tanuló modellek segítségével valósul meg a predikció. Az elvégzett adatelemzési feladat komplexitását leíró táblázat az alábbiak szerint alakult:

Termék értékelések elemzése és vizualizálása:
Ennek kapcsán egy olyan PILOT feladaton dolgoztunk, amely célul tűzte ki az értékesített termékekre adott visszajelzések elemzését, hogy a vásárlók (akár az érékesítők is) egyszerűen, áttekinthető módon be tudják azonosítani a termékekhez kapcsolódó legfontosabb pozitívumokat és negatívumokat. Megoldásunk alapját természetes nyelvi adatok feloldozásához fejlesztett eszközök adták, kiegészítve újszerű megoldásokkal és vizualizációs eszközökkel.

A megoldás felhasználható akár weblapokhoz, beépülő modul formájában, amely megjelenik a termékek alatt található értékelések előtt és egyszerűen, áttekinthető módon megmutatja azok lényegét, valahogy így:



A fejlesztés során olyan egyedi kulcsszó kinyerő algoritmus fejlesztésére volt szükség, amely képes rövid, akár néhány mondatos szövegből kinyerni azokat a jellemző kulcsfrázisokat, amelyeket érdemes megjeleníteni a felhasználó számára. Az elvégzett adatelemzési feladat komplexitását leíró táblázat az alábbiak szerint alakult:

8.4. Példa 4: Egyedi elemzés

9. TEVÉKENYSÉGMEGHATÁROZÁS AZ ADATELEMZÉS JELLEGE ALAPJÁN

A feladat komplexitása alapján előre megbecsülhető, hogy az adott feladat elvégzéséhez (a tapasztalat alapján) milyen az időszükséglet. A pontos becslés elkészítéséhez azonban szükséges meghatározni az elemzési feladat jellegét is, melyből a feladat végrehajtásához szükséges tevékenységek pontos meghatározása történik.

	Leíró jellegű elemzés/riport vagy megfigyelés	Feltáró elemzés	Mély elemzés prediktív analitikával kiegészítve	Egyedi elemzés, tanulmány
I.a/b Konzultáció az előzőek alapján	Igen	Igen	Igen	Igen
I.c Adatelemzési feladat megfogalmazása	Igen	Igen	Igen	Igen
I.d Az adatelemzési terv elkészítése	Igen	Igen	Igen	Igen
II.a Adatok összegyűjtése	Igen	Igen	Igen	Igen
II.b Adatok leírása	Igen	Igen	Igen	Igen
II.c Adatok feltárása		Igen	Igen	Igen
II.d Az adatminőség ellenőrzése			Igen	Igen
III.a Adatok kiválasztása			Igen	Igen
III.b Adattisztítás		Igen	Igen	Igen
III.c Adatszármaztatás			Igen	Igen
III.d Adatok integrálása		Igen	Igen	Igen
III.e Adatok formátumának módosítása			Igen	Igen
IV.a Modell kiválasztása	Igen	Igen	Igen	Igen
IV.b Tesztelési módszertan kialakítása			Igen	Igen
IV.c Modell építése a rendelkezésre álló adathalmaz segítségével	Igen	Igen	Igen	Igen
IV.d Modell kiértékelése			Igen	Igen
V.a Eredmények kiértékelése	Igen	Igen	Igen	Igen
V.b Alkalmazott módszertan áttekintése	Igen	Igen	Igen	Igen
V.c Következő lépések meghatározása			Igen	Igen
VI.a Eredmények megjelenítése	Igen	Igen	Igen	Igen
VI.b Zárójelentés készítése	Igen	Igen	Igen	Igen

Természetesen minden elemzési feladat egyedi, így az elemzéshez elvégzendő, szükséges tevékenységek is változhatnak.

10. ÖSSZEFOGLALÁS

A fentiekben vázolt modell egy ajánlott adatelemzési folyamatot mutat be, valamint egy arra épülő feladatmátrixot, azaz az egyes lépések szerepét a különböző típusú és komplexitású adatelemzési feladatok esetén. A cikkben nem foglalkoztunk bővebben a bevezetés lépéseivel, valamint az azt követő üzemeltetési feladatokkal és kihívásokkal.

Jegyzetek

1 Wirth, R., Hipp, J. (2000) CRISP-DM: Towards a Standard Process Model for Data Mining, In: *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29-39. http://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf (2021. 09. 28.).

2 Provost, F., Fawcett, T. (2013) Data Science and its Relationship to Big Data and Data-Driven Decision Making, *Big Data*, 1(1), 51-59., https://doi.org/10.1089/big.2013.1508 (2021. 09. 28.).

3 Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000) CRISP-DM 1.0., https://the-modeling-agency.com/crisp-dm.pdf (2021. 09. 28.), 76.

4 Angée, S., Lozano-Argel, S. I., Montoya-Munera, E. N., Ospina-Arango, J-D., Tabares-Betancur, M. S. (2018) Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects, In: Uden, L., Hadzima, B., I-Hsien Ting (eds.) *Knowledge Management in Organizations* (Cham: Springer) 613-624., https://doi.org/10.1007/978-3-319-95204-8_51 (2021. 09. 28.).