

Budapesti Corvinus Egyetem

Gazdálkodástudományi Kar

Információrendszerek Tanszék

Oktatási adatbányászat és analitika alkalmazása középiskolai adatokon

Készítette: Sepp Ákos Norbert

gazdaságinformatikus alapszak

2018

Szakszeminárium-vezető: Dr. Vas Réka Franciska

Tartalomjegyzék

Tartalomjegyzék	2
1 Bevezetés.....	4
2 Az oktatási analitika	5
2.1 Oktatási adatelemzések főbb területei	5
2.2 Az oktatási adatbányászat	6
2.3 Az osztályzáselemzés (Grading Analytics)	8
2.4 A CRISP-DM módszertan	9
2.5 A CRISP-DM módszertan alkalmazása a dolgozatban	11
3 Az oktatás-informatikai környezet	13
3.1 Információs társadalom az oktatásban	13
3.2 Magyarország oktatás-informatikai helyzete	14
3.3 Az e-napló	15
4 A tanulmányi és érettségi adatok közötti kapcsolat vizsgálata	17
5 Tanulmányi és érettségi adatok	19
5.1 Nyers adatok bemutatása	19
5.2 Az elemzéshez szükséges adatforma	20
5.3 Az adatok elemzésre vetített korlátai	21
6 Az adatelőkészítés folyamata	23
6.1 Irreleváns adatok eltávolítása.....	23
6.2 Egyes attribútumok elemzésre alkalmassá tétele	24
6.3 Az érettségi eredmények objektív összehasonlítása	25
6.4 Az attribútumok aggregálása	27
6.5 Az adatkörök illesztési feltételeinek megteremtése	29

6.6	Az adatkörök illesztése, adatok anonimizálása.....	29
6.7	Egy adatelőkészítési javaslat a jövőbeli elemzésekhez	30
7	Osztályzáselemzés prediktív módszerrel	31
7.1	Az alkalmazott modell típusok bemutatása	31
7.2	Modellezési folyamat.....	32
7.3	Választás a prediktív modellek között	34
7.4	Modell felépítése, kapott eredmény bemutatása	35
8	A prediktív modell értékelése	38
8.1	Mutatószerzők szerepe az értékelésben	38
8.2	A prediktív modell eredményének értékelése.....	40
8.3	Tanulmányi eredmények szerepének újraértékelése klaszterezéssel.....	41
8.4	Az osztályzás elemzés alkalmazási lehetőségei az intézményben.....	46
9	További osztályzáselemzési lehetőségek	47
9.1	A jelenlegi modell továbbfejlesztése	47
9.2	Újabb adatkörök bevonása az osztályzáselemzésbe	48
10	Összefoglalás	49
	Irodalomjegyzék	51
	Ábragyegyzék.....	54

Köszönetnyilvánítás

Szeretném kifejezni köszönetemet, amiért a Prohászka Ottokár Katolikus Gimnázium a rendelkezésemre bocsátotta a dolgozat szempontjából elengedhetetlen adatokat.

Köszönet illeti konzulensemét, Dr. Vas Réka Franciskát. Munkájával, mindig építő szándékú kritikáival és tanácsaival nagy mértékben hozzájárult, hogy ez a dolgozat elkészülhetett.

Végezetül köszönöm édesapámnak, Sepp Norbertnek minden segítségét, ötletét és meglátását.

1 Bevezetés

„Mikor megpróbáljuk elképzelni a jövő oktatását, a hangsúly általában az új technológián van, a mindenhol jelen lévő számítógépeken, a rugalmas tanterem modelleken, és az innovatív vizuális kijelzőkön. Pedig a jövő felsőoktatásának alakításában a legmeghatározóbb összetevőt valami olyan jelenti, ami szemmel nem látható, és kézzel nem megfogható: a Big Data és az analitika.” (Long és Siemens, 2011, 1. old.)

Az oktatás reformja olyan kérdés, ami mindenhol jelen van. Habár nincs egyetértés azzal kapcsolatosan, hogy mi minden szükséges az átalakulás véghez viteléhez, az oktatás analitikai megújulását kiemelkedően fontosnak tartják a területtel foglalkozó kutatók. Ahogy tanulmányukban megfogalmazzák, „a tanulóelemzés a fejlődés ígérését hordozza a tanulás hatékonyságának és sikerességének növelésében az általános iskolák, középiskolák és a felsőoktatás számára.” (Siemens et al, 2011, 5. old.)

Baker és Siemens (2014) kiemelik, hogy a tanulást kutató tudományok komoly, néhol több évtizedes késésben vannak az analitika alkalmazását tekintve. Ahogy az az elmúlt évek kis lépésekben történő, sokszor szigetszerű változásait tekintve megállapítható, az oktatás analitikai reformja alulról szerveződő (bottom-up) módon megy végbe.

Az oktatási rendszerek nagyon nagy különbségeket mutatnak az egyes intézményi területeken és szinteken, gondoljunk csak a különböző földrajzi régiók technológiai szakadékaiba, vagy a különböző tudományok oktatására szakosodott intézmények közti eltérésekre. Ebből kifolyólag egy egységesített, felülről jövő (top-down) megoldás nem lenne se hatékony, se célravezető. Azonban úgy gondolom, minden tudományos munka, amely a területen megszületik, akár kisebb elemzések, akár nagyobb volumenű kutatások keretében, előreviszi az oktatás régóta szükséges reformját. Ebből a megfontolásból választottam én is dolgozatomban témájaként az oktatással kapcsolatos adatelemzést.

Dolgozatomban egy szűkebb értelmezési tartományt, a középiskolai környezetet vettem alapul, ezen belül is egy konkrét középiskolai intézmény adatait elemeztem. Céлом az oktatási adatelemzést alkalmazva, konkrét kérdéseken keresztül, többlet tudás nyújtása az adott intézmény számára, valamint egy, az elemzés eredményének gyakorlatba ültetésére adott javaslat elkészítése.

2 Az oktatási analitika

2.1 Oktatási adatelemzések főbb területei

Elsőként a területhez kapcsolódó fogalmakat szeretném definiálni. Long és Siemens (2011) alapvetően két fő típusra bontja az oktatással kapcsolatos elemzéseket: tanulóelemzésre és akadémiaelemzésre. Az 1. táblázatban foglalom össze az alapvető különbségeket a két típus között.

A tanulóelemzés (Learning Analytics, LA) az adat mérése, gyűjtése, elemzése és riportolása a tanulókra és környezetükre vonatkozóan, a tanulás folyamatának valamint a hozzá kapcsolódó környezetnek a megértése és optimalizálása céljából. A tanulóelemzés nagyrészt tehát a tanulók sikerességének növelését szolgálja.

Az akadémiaelemzés (Academic Analytics, AA) célja a szervezeti folyamatok, workflow-k, erőforrás-allokációk fejlesztése. Összességében tehát az oktatásnak helyt adó intézmény vagy szervezet hatékonyságának növelését szolgálja.

Elemzés típusa	Elemzés szintje / tárgya	Ki profitál az elemzésből?
Tanulóelemzés	Kurzus-szint: kapcsolati hálók, fogalmi fejlődés, társalgás-elemzés, intelligens tanterv	tanulók, intézmény
	Intézményes: előrejelző modellezés, siker/kudarcminták	tanulók, intézmény
Akadémiaelemzés	Szervezeti: tanulói profilok, tanulmányi eredmények, tudásfolyam	adminisztrátorok, befektetők, marketingesek
	Regionális: rendszere közti összehasonlítás	befektetők, adminisztrátorok
	Nemzeti és nemzetközi	nemzeti kormányok, oktatásért felelős szervek

1. táblázat: Tanulóelemzés és akadémiaelemzés (Forrás: Baker és Siemens, 2014, 34. oldal)

Azt is megfogalmazzák kutatásukban, hogy az oktatási adatelemzések területe nagy hasonlóságot mutat az üzleti intelligenciával, így itt is fontos, hogy egyfajta fejlődési ciklust lehessen felállítani az egyes adatelemzési szintek között. Az alábbi sorrendet határozták meg:

- 1) **Kurzus-szintű adatelemzés:** tanulási utak, kapcsolati hálók, kommunikációs folyamatok elemzése
- 2) **Oktatási adatbányászat:** előrejelző modellezés, klaszterezés, mintakeresés
- 3) **Intelligens tanterv:** tartalom alapján meghatározott tantervi erőforrások
- 4) **Adaptálható tartalom:** tartalomalapú, tanulói viselkedésen alapuló ajánlórendszer
- 5) **Adaptálható tanulás:** nemcsak a tartalom, hanem az egész tanulási folyamat adaptálható az egyes tanulókra

Az öt szint közül dolgozatomban az oktatási adatbányászatot választottam alkalmazott adatelemzési szintnek.

2.2 Az oktatási adatbányászat

Magát az adatbányászatot Witten és Frank (1999) úgy határozta meg, mint „az a terület, amely újszerű és potenciálisan hasznos információt fedez fel nagy mennyiségű adatból.” (Baker és Yacef, 2009).

Ezzel párhuzamban „az oktatási adatbányászat egy feltörekvő ágazat, amely az oktatáshoz köthető környezet egyedi adataival kapcsolatban hoz létre feltáró módszereket, és alkalmazza ezeket a diákok és környezetünk jobb megértésének céljából” (Baker és Yacef, 2009, 4. old.).

A két megfogalmazás, valamint a terület beszédes neve is jól mutatja, hogy az oktatási adatbányászat egy olyan részterülete az adatbányászatnak, melynek központjában a diákok, valamint környezetük áll. Ugyanakkor, ahogy Baker (2010) kiemeli, az oktatási adatbányászat módszerei sokszor különböznek a hétköznapi adatbányászat által alkalmazottaktól, egyszerűen azért, mert explicit adatokra van szükség a többszintű hierarchia és a nem-független oktatási adatok miatt.

Baker és Yacef (2009) erre vezeti vissza, hogy az oktatási adatbányászat módszerei különböző szakterületekről állnak össze, ideértve az adatbányászatot és a gépi tanulást, a pszichometriát és más statisztikai területeket, az informatika adat-vizualizációt és a számítógépes modellezést.

Baker és Siemens (2014) az alábbi módon kategorizálja az oktatási adatbányászat fő módszereit és eszközeit:

Előrejelzés

A cél, hogy egy olyan modellt fejlesszünk, amely egyetlen adattényezőre (statisztikából is ismert függő változó, vagy célváltozó) enged következtetni az adathalmaz többi részének bizonyos kombinációjából. Az előrejelzésnek három fő típusát szokták alkalmazni: a *osztályozást*, a *regressziót* és a *rejtett tudás becslést*.

Az *osztályozó* modellnél a célváltozó lehet bináris (pl. 0 vagy 1), vagy kategóriai változó. A leggyakoribb eszközei közé tartozik a döntési fa vagy a véletlen erdő. A *regresszió*nál a célváltozó egy folytonos érték, például egy szám. A leggyakoribb példa rá a lineáris regresszió. A harmadik típus a *rejtett tudás becslés*. Ez a tanulók tudását vagy speciális képességeit és koncepcióit veti össze a kritikus képességek helyességével.

Struktúra felfedező algoritmusok

Az előrejelzéssel szemben itt nincs egy előzetes elképzelés azzal kapcsolatban, mit kellene találni az adatokban, így nincs célváltozó sem. A kutatónak itt pontosan az a célja, hogy meghatározza, milyen struktúra rejlik az adathalmazban. A leggyakoribb megközelítések a *klaszterezés*, a *faktoranalízis*, a *kapcsolati háló elemzés* és a *fő struktúra felfedezés*.

A *klaszterezés* célja, hogy az adatokban önállóan kialakuló csoportokat figyeljünk meg. Ha jól határoztuk meg ezeket a klasztereket, minden adat alapvetően több hasonlóságot mutat bármely klaszterbeli adattal, mint más klaszterbeli adatokkal. A *faktoranalízis* egy szorosan kapcsolódó módszer, szintén azt célozza, hogy egymáshoz közelálló adatszoportokat határozzunk meg, mégpedig rejtett változók mentén. A faktoranalízist gyakran alkalmazzák, amikor nagyszámú változóból szeretnék kiszűrni a lényeges dimenziókat. A *kapcsolati háló elemzés* a kapcsolatot és interakciókat vizsgálja az egyes szereplők között, valamint az ebből keletkező adatokat. A *fő struktúra felfedezés* célja pedig a tudás struktúráját megtalálni egy oktatási területnek, például milyen kapcsolat fedezhető fel a jellemző tudás vagy képességek és a jellemző tartalom között.

Kapcsolatbányászat

Célja az adathalmaz egyes változói között megtalálni a kapcsolatot. Alapvetően négy típusa van, az *asszociációs szabály bányászat*, a *korrelációbányászat*, az *ismétlődő mintázat bányászat* és az *alkalmi adatbányászat*.

Asszociációs szabály bányászat az a folyamat, amikor ha-akkor szabályokat keresünk egyes változók között. *Korrelációbányászat* a folyamat, ha a célunk bizonyos változók közötti lineáris korrelációs kapcsolatot találni. Az *ismétlődő mintázat bányászat* folyamata során időszakos összefüggéseket keresünk az események között. Az *alkalmi adatbányászat* során arra a kérdésre keressük a választ, hogy egy adott esemény vajon kiváltója volt-e egy másik eseménynek. Ami mind a négy módszerben közös, hogy potenciálisan váratlan, de jelentős kapcsolatokat fedezhetünk fel egyes változók között, amiket aztán szélesebb körben is használhatunk, akár új elméletek generálására, akár automatikus előrejelző rendszerek lehetséges közbeavatkozási szabályainak meghatározására.

Adatvizualizáció az emberi megítélés számára

Célja, hogy a gyors döntéshozáshoz és az azonnali reakcióhoz szükséges lényeges adatokat megjelenítse. Többek között olyan módszerek említhetők itt meg, mint a *heat map*, a *tanulási görbe*, vagy a *learnogram* (mely a tanuló tevékenységei közötti váltásokat ábrázolja, ahogy azt Herschkovitz és Nachmias, 2008 definiálják).

Modellekkel történő felfedezés

Egy korábban felállított modell újrafelhasználása egy másik elemzésben. Az egyik leggyakoribb alkalmazása, amikor egy előrejelző modellt alkalmaznak egy másik előrejelző modell részeként, például egy függő változóként. A másik népszerű alkalmazási formája, ha egy előrejelző modellt illesztnek egy kapcsolatbányászati modellbe. Ilyenkor a korábbi célváltozó és a további változók közötti kapcsolatot tudják elemezni, ami lehetőséget nyújt a kutatók számára egy széles adathalmaz és egy összetett, rejtett felépítés közötti összefüggések tanulmányozására.

2.3 Az osztályzáselemzés (Grading Analytics)

Az oktatási adatbányászat jellemző módon az elemzéshez szükséges adatokat a tanulók és oktatók által hagyott digitális lábnyomokból nyeri ki, például egy LMS (Learning Management System) vagy OLT (Online Learning Tool) rendszeren keresztül, vagy esetleg egy osztálytermi vagy más intézményi megfigyelő rendszeren át. Ugyanakkor itt két komoly probléma is felmerül az alkalmazáskor. Egyfelől, nem minden oktatási intézmény engedheti meg magának, hogy digitális megfigyelést és nyomkövetést alkalmazzon a tanulás és oktatás folyamata során,

másfelől egy ilyen szintű megfigyelés több helyen is törvényi korlátozásokba ütközik, a személyiségi jogok védelme miatt (Zielke et al., 2018).

Zielke és társai (2018) egy olyan megközelítést alkalmaztak az adatbányászatra, ami úgy nyújt adatelemzési lehetőséget, hogy nem a tanulási és oktatási folyamatból kinyert adatokra épül, hanem azokkal az eszközökkel is megvalósítható, amik egy átlagos oktatási környezetben megtalálhatóak. A legfontosabb adat szerepét így a tanulók osztályzatai és vizsgaeredményei jelentették. Ezt a megközelítést nevezték el osztályzáselemzésnek (Grading Analytics).

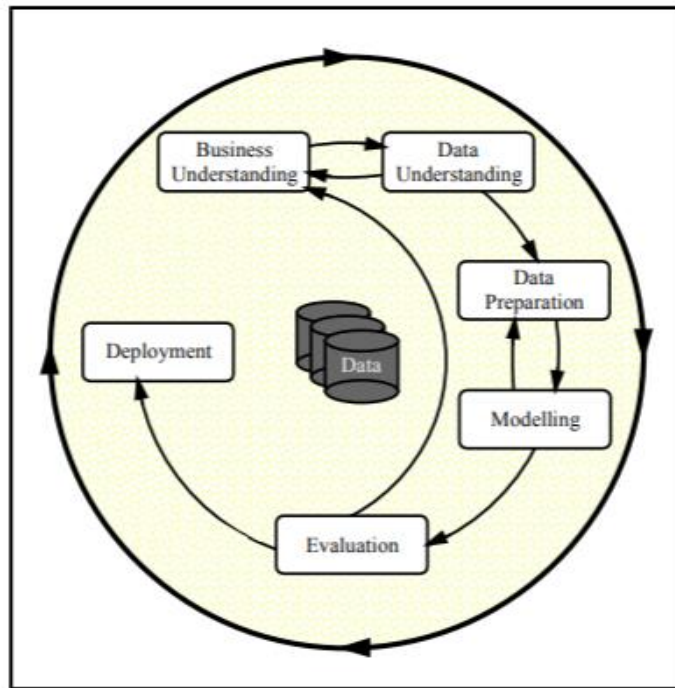
Kutatásukban a tanulók hosszú évre visszamenőleg rögzített, valós tanulmányi eredményeiből állítottak fel különböző előrejelző modelleket, amelyek segítségével megmutatták, hogy egy szak elvégzésének végeredményét sikeresen meg lehet becsülni, akár a nyers adatok (egy-egy tanulók konkrét évközi eredményei) közvetlen felhasználása nélkül is.

Dolgozatomban egy olyan középiskolai intézmény adatait vizsgálom, ahol az oktatási adatbányászat osztályzáselemzési ága az, amit meg lehet valósítani. A rendelkezésre álló kiindulási (nyers) adatok nagyjából megegyeznek Zielke és társai kutatási helyzetével, ugyanakkor dolgozatomban az értelmezési környezet a középiskola, nem a felsőoktatás.

2.4 A CRISP-DM módszertan

Az adatbányászat egy kreatív folyamat, ahol különböző készségek és tudás szükséges a cél eléréséhez. Sokáig nem volt egységes keretrendszere az adatbányász projekteknek, a siker jórészt azon állt, hogy a kutató vagy kutatócsapat hogyan végezte a munkát, ez pedig magában hordozta azt is, hogy a módszerek nehezen voltak megismételhetőek. A CRISP-DM (Cross Industry Standard Process for Data Mining) projekt célja az volt, hogy ezeket a problémákat felismerve egy olyan modellt dolgozzanak ki, ami függetlenül az iparágtól és az alkalmazott technológiától, keretrendszert biztosít az adatbányász projekteknek, foglalja össze Wirth és Hipp (2000).

A CRISP-DM egy adatbányászati életciklust fogalmaz meg, ami a projekt fő lépéseit tartalmazza (1. ábra).



1. ábra: A CRISP-DM életciklus (Forrás: Wirth és Hipp, 2000, 33.oldal)

Az egyes lépésekhez a következő jelentéstartalmat társítják:

Üzleti helyzet megértése

Az első lépés a projekt céljainak és követelményeinek üzleti szempontból történő megértésére fókuszál, majd ezt a tudást fordítja le egy adatbányászati problémára, valamint egy elsődleges projekttervet állít fel a célok eléréséhez.

Adatok megértése

Ez a lépés az adatok begyűjtésével és azok megismerésével foglalkozik. Itt történik az adatok minőségbeli problémájának felismerése, az első benyomások az adatokkal kapcsolatban ekkor születnek, és ilyenkor kerülnek felszínre érdekes adatrészek is, amelyek hipotézisek felállításához vezetnek.

Adatelőkészítés

Ebben a lépésben történik minden folyamat, amely a végleges adathalmaz előállítását szolgálja. Ide tartozik a teljesség igénye nélkül az adatok rendezése, többszörözése, tisztítása és átalakítása.

Modellezés

Ez a lépés különböző modellező technikák kiválasztását és alkalmazását foglalja magában. Jellemzően több elemzési technika is létezik egyazon adatbányászati problémára. Ez a lépés szoros kapcsolatban áll az adatelőkészítési fázissal, mert gyakran csak ilyenkor kerülnek felszínre az adatok hiányosságai, vagy ilyenkor fogalmazódik meg egy új ötlet az adatstruktúrához.

Kiértékelés

A projekt e pontján már egy vagy több kellően magas minőségű modell készült az adatelemzéshez. Mielőtt azonban a modellt véglegesítenénk és élesben alkalmaznánk, fontos hogy alaposabban megvizsgáljuk és kiértékeljük, vajon a modellt valóban az eredeti céloknak megfelelően állítottuk-e fel. A lényeges szempont az, hogy van-e olyan fontos üzleti probléma, ami nem lett megvizsgálva. A kiértékelés lépés végén jutunk el a döntésig, hogy alkalmazzuk-e a modellt vagy sem.

Alkalmazás

A modell elkészítése nem a projekt utolsó lépése. Az összegyűjtött tudást rendszerezni és prezentálni kell olyan módon, hogy a felhasználó vagy megrendelő azt értelmezni tudja. A kiinduló követelményektől függően ez a lépés állhat egy egyszerű riportból, de nehezebb esetekben akár egy megismételhető adatbányászati folyamat felállítását is jelentheti. Sok esetben akár nem is az adatelemző, hanem a felhasználó az, akinek ezt a lépést végre kell hajtania. A lényeg, hogy meg kell érteni, pontosan milyen tevékenységek által lesz valódi haszna az elkészült modelleknek.

2.5 A CRISP-DM módszertan alkalmazása a dolgozatban

Dolgozatomat a CRISP-DM módszertan alapján kívánom felépíteni. Az egyes lépéseknek megfelelően elsőként azt a környezetet elemzem, amelyre a kutatás témája fókuszál, valamint ahonnan az adatok is származnak, vagyis a magyarországi középiskolai informatikai rendszereket. A környezet **megértése** során történik az üzleti igény átfordítása adatbányászati problémára, vagyis ekkor fogalmazódnak meg a dolgozat főbb kérdései.

Második lépésként ismertetem a rendelkezésre álló nyers **adatok** formáját, tartalmukat, mennyiség- és minőségbeli jellemzőiket. Bemutatom a középiskolai intézményhez tartozó adatkörökből kinyert adathalmazokat.

A harmadik lépés az adatok **tisztítása**. Az elemzéshez szükséges formát, valamint az oda elvezető út egyes fázisait mutatom be ábrák és leírások segítségével. Ez az előkészítési lépés tartalmazza azt az adattárolási és kinyerési folyamat-javaslatot is, ami a jövőbeni, hasonló témájú kutatások számára már előrelépést jelenthet.

Az adatelőkészítés után a **modellezés** bemutatása következik. Több modell alkalmazását is felvetve, az egyes modellek közötti eltéréseket számításokkal alátámasztva végzem a kiválasztást, majd a megfelelő modellt alkalmazom a már megtisztított adathalmazon.

A modell felállítását követi a **kiértékelés** lépése. A szakirodalom által ajánlott mutatószámok mentén, valamint tanult statisztikai módszerekkel értékelem a kidolgozott modell eredményeit. Mindezek fényében pedig elfogadásra, elutasításra vagy feltételes elfogadásra kerül a modell.

Végül az **alkalmazási** lépésben megfogalmazok egy olyan módszert, amellyel a gyakorlatban is hasznosítható a kidolgozott modell, egy konkrét példán keresztül bemutatva. Ezután egy javaslatot teszek, ami egyfelől belépési pontokra koncentrálna a modell alkalmazásának lehetséges folyamatába, másfelől arra, hogyan használható a kimeneti tartalom a választott intézmény hétköznapi gyakorlatában.

3 Az oktatás-informatikai környezet

3.1 Információs társadalom az oktatásban

A 21. századi technológia jelentős mértékben átalakította a társadalmat, a társadalmi tőkét. A ma kulcsfontosságúnak tekinthető készségek és képességek, melyek közül a legtöbb az IKT-s (információs-kommunikációs technológia) eszközökhöz kapcsolódik, hatékony használata nélkül elképzelhetetlen a jelen és a jövő tanítási környezete. Ennek kialakulását elősegíti a társadalmi környezet, ahol közösen szerkesztünk, készítünk és osztunk meg egymás közt médiatartalmakat, tartjuk a kapcsolatot online a hétköznapiak mindegyikén. Ez a fajta, átalakuló kompetencia, képesség és tudáskör pedig új kihívás elé állítja az oktatást.

Molnár (2011) szerint ennek fontosságát jól jelzi, hogy az USA és az EU is célként fogalmazta meg és támogatja az IKT-használattal összefüggő készségek, kompetenciák oktatásba történő integrálását. Az EU 2010-ig elérendő, IKT-t érintő oktatáspolitikai ajánlásában többek között ezeket a kritériumokat fogalmazta meg:

- az IKT-s eszközök jelenjenek meg az oktatás teljes területén, a képzés ne korlátozódjon a felhasználói ismeretek átadására
- az informatikai eszközök használata készségszinten épüljön be a tanítás-tanulás folyamatába
- a tananyag tartalma és hozzáférhetősége legyen rugalmasabb, nyitott tanulási környezetben történjen az oktatás
- az iskola alakítsa ki, és fejlessze az élethosszig tartó tanuláshoz szükséges alapkészségeket, s készítsen fel az új tanulási formák használatára.

Hozzáteszi, szintén nagy probléma, hogy annak ellenére, hogy a technológia, bár különböző szinten és elterjedtséggel, de közel 30 év óta az oktatási rendszerek részét képezik, nem fogalmazódott meg egységesen, nemzetközi szinten olyan indikátor rendszer, amely konstruktívan és objektíven összehasonlíthatóvá tenné, mennyire sikerült integrálni a technológiát az oktatás folyamatába.

Molnár munkájából kiderül, hogy a nyugati világ egyetért abban a célkitűzésben, hogy az oktatásnak nyitnia kell a modern technológia integrálása felé.

Az oktatás megvalósulásának szintjén, az egyes tanórákon és intézményekben tehát kiemelt fontosságúnak ítélték meg a modern technológia integrálást. De mi a helyzet az eggyel magasabb szinttel? Mi történik azokkal az adatokkal, amik egy-egy ilyen alkalmazási folyamat során jönnek létre?

Talán nem túlzás kijelenteni, hogy ha az oktatás szintjén kiemelten fontosnak tartjuk a technológiai fejlődés, akkor ugyanezt elvárhatjuk az oktatás hatékonyságának, eredményességének mérésekor is, hiszen e mutatók a visszacsatolás és értékelés legfontosabb kritériumai, melyekkel meg tudjuk ítélni az egész folyamat sikerességét.

Fontos megvizsgálnunk azt is, milyen lépések történtek a világban és Magyarországon azóta, hogy a fent megfogalmazott célok ki lettek tűzve.

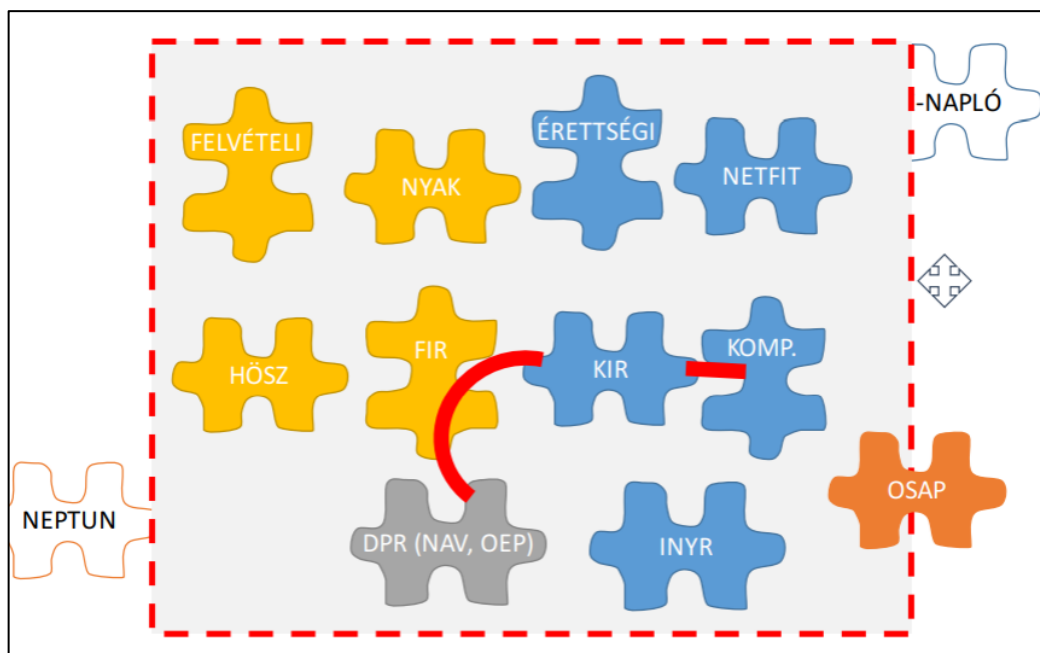
3.2 Magyarország oktatás-informatikai helyzete

Fodor és társai (2016) az alábbi táblázatban foglalták össze a globális és hazai viszonyokat az oktatás digitális helyzetét tekintve (2. táblázat):

Globális helyzet	Hazai helyzet
- kartotékok helyett elektronikus nyilvántartás: adatok robbanásszerű növekedése - kormányzati, gazdasági és társadalmi elvárások az oktatással szemben: eredményesség, mérhetőség, elszámolhatóság - nemzetközi jó gyakorlatok: Anglia, Hollandia, Svédország - adatvédelmi kérdések - nyilvánosság	- szigetszerű fejlesztések - adatvagyon kihasználtsága gyenge, nem épül be a folyamatokba (csak a köznevelésben több mint 3000 adatkör, 168 elkülönült adatbázis létezik) - ad hoc és nem sztenderdizált kapcsolat az adatbázisok között: pontatlanság, redundancia - jó gyakorlat: kompetencia-adatbázis

2. táblázat: Globális és hazai digitális oktatási viszonyok (Forrás: Fodor és társai, 2016, 3. old.)

Ahogy az a táblázatból kiolvasható, 168 elkülönült adatbázis van jelenleg jelen a magyar oktatásban, amelyek egyenként olyan fontos adatokat tárolnak, mint például az érettségi eredmények, a kompetencia mérések, a NETFIT, a KIR, az OKTV, a felvételi eredmények, vagy az e-napló. Ahogy azt Fodor és társai kiemelik (2. ábra), az egyes adatbázisok között azonban nincs, vagy nagyon alacsony szintű az integráció.



2. ábra: Oktatással kapcsolatos adatbázisok Magyarországon (Fodor és társai, 2016, 5. old.)

A dolgozathoz szükséges adatgyűjtés során tapasztaltam, hogy ezek az adatkörök nem csak az integráció szintjét nem érik el, de előfordult, hogy digitalizálva sincsenek.

3.3 Az e-napló

Funkciója

A szoftver hivatalos weboldalának megfogalmazása szerint „a közösség erejével fejlesztett és karbantartott MaYoR elektronikus osztálynapló szoftvert a mindennapi adminisztrációs feladatok hatékony szervezésére, terhének elosztására és a bonyolult központi szabályrendszerek egyszerű betartására tervezték, figyelembe véve az iskolák szereplőinek együttes igényeit, és az egyedi szokásoknak megfeleltethető konfigurálhatóságot is.” (MaYoR, 2018)

Technológiai háttere

A PHP alapú webes alkalmazás keretrendszere MySQL, MS Active Directory és LDAP back-endből tud felhasználókat azonosítani. A szoftver platform-független, GPLv2 licenszelésű ingyenes szabadsoftver.

A szoftver célja elsősorban a papíralapú naplózás digitálisra való áttérésének elősegítése az egyes intézményekben. Ennek megfelelően szolgáltatásai körébe olyan funkciók tartoznak, amelyek a papíralapú adminisztrációnak digitális megfelelői. A teljesség igénye nélkül: felhasználók adatainak tárolása, éves munkaterv és időszakok kezelése, tankörök és osztályok nyilvántartása, haladási napló és kapcsolódó megjegyzések rögzítése, jegyek és dolgozatok, bizonyítványok és tanulmányi előmenetek kezelése, hiányzások és órarendek adminisztrálása.

Működése a választott intézményben

A dolgozat adatait szolgáltató intézmény a MaYoR e-naplót alkalmazza az oktatás mindennapjaiban. Az intézményben 2013 óta a szoftver szolgáltatásainak köszönhetően az adminisztráció majdnem teljes része itt történik.

A rendszer a tanulók adatainak nyilvántartásán túl az órarendek összeállításának és rugalmas, ideiglenes változtatásainak követését szolgálja, valamint osztályozónaplóként is funkcionál. A hiányzások, kikérők, visszajelzések adminisztratív rendszereként is kapcsolatot tart a pedagógusok és a tanulók szülei között.

4 A tanulmányi és érettségi adatok közötti kapcsolat vizsgálata

A dolgozat során két, a magyar oktatásban széleskörben jelenlevő adatkört illesztettem össze, az oktatási adminisztrációt tartalmazó e-naplót, valamint a középiskola elvégzésének eredményeit rögzítő érettségi eredményeket. Ez a kettő adatkör azt írja le, hogy egy adott tanuló milyen eredményekkel vette az oktatási folyamat mérföldköveinek tekinthető féléves és évvégi akadályokat, valamint hogy mekkora sikerrel teljesítette az érettségi vizsgát.

Dolgozatom kérdéseinek alapja a tanulmányi és az érettségi eredmények közötti kapcsolat vizsgálata. Mindkét adatkör vizsgálata nagyon érdekes, akár önmagukban is fontos elemzési témát nyújthatnak, összekapcsolásuktól azonban többlet tudás kinyerését várom, akár ok-okozati szinten is.

Az elemzéshez rendelkezésre álltak az érettségi adatok a 2011-2016. év közötti időszakot felölelően, és az e-naplóból kinyert adatok a 2012-2016. évi időszakra vonatkozóan. A két különböző adatkör sajátosságaiból fakadóan az előbbi mindig a végzős, valamint előrehozott érettségit tevő diákokról rögzít adatokat, utóbbi viszont a teljes diákságról (3. táblázat).

tanév / évf.	5	6	7	8	9	10	11	12	érettségi adatok (x)
2011	nincsenek adatok							x	
2012	.	.	.	.	.	.	.	x	
2013	.	.	.	.	.	.	.	x	
2014	.	.	.	.	.	.	.	x	
2015	.	.	.	.	.	.	.	x	
2016	.	.	.	.	.	.	.	x	
2017	.	.	.	.	.	.	.	n.a.	
e-napló adatok (.)									
Jelmagyarázat:			.	= az adott osztályról nincsen érettségi / e-napló adat					
			.	= az adott osztályról van érettségi adat					
			x	=az adott osztályról az évben érettségi adat is van					

3. táblázat: Rendelkezésre álló adatállományok (Forrás: saját szerkesztés)

Az e-naplóból, a bizonyítványban szereplő érdemjegyeken túl, rendelkezésre áll néhány kiegészítő adat, amelyek árnyalhatják az osztályzáselemzés képét: a diák neve, életkora, félév során összegyűjtött hiányzásai, magatartása és szorgalma. Ezeket az adatokat a dolgozatban „többlet adatoknak” nevezem, melyek a következtetések képét árnyalhatják, valamint további vizsgálatokra vonatkozó lehetőségeket rejtenek magukban.

A tanulmányi eredmények egyszerre fejezik ki számszerűen és összehasonlítható módon azt a sokéves teljesítményt, amit a tanuló a középiskolai oktatásban nyújt, illetve számítanak bele a felsőoktatási felvételi pontszámaiba. Ez azt jelenti, hogy nem csak a tizenkét éves oktatásban eltöltött teljesítményt értékelik, hanem hosszú távú hatásai is vannak a tanuló továbbtanulási lehetőségeire.

Az érettségi eredmények pedig egy másik szempontból, de hasonló módon nyújtanak tükröt az általános és középiskolai évek során elsajátított tanulói tudásról, ennek köszönhetően jelentenek sok esetben akár munkaképes végzettséget is. Ezen túl pedig 80%-ban meghatározzák egy-egy tanuló továbbtanulási esélyeit (az összesen gyűjthető 500 pontból 400-at az érettségi pontok tesznek ki).

A dolgozat fő kérdését így fogalmaztam meg: milyen összefüggés mutatható ki a tanulmányi és érettségi eredmények között, és ezt az ismeretet hogyan lehet az oktatásban közvetlenül kamatoztatni?

A tervezett válasz első része egy olyan modell felállítása, amely segítségével elfogadhatóan jó közelítést adhatunk az előbbi adatkör adataiból az utóbbi eredményére. A modell felállításához prediktív módszert használva, előrejelző modellt építünk. A modell célja egyfajta becslést nyújtani a tanulmányi adatokból a várható érettségi eredményekre.

A válasz második része a modell gyakorlati alkalmazási lehetőségeinek vizsgálata. A modell gyakorlatba ültetését egy konkrét példán keresztül vizsgálom, ami a tanulmányi eredmények érettségi eredmények szempontjából betöltött szerepének sztereotip kérdését vizsgálja.

A dolgozat részeként végül egy rövid javaslatot is teszek a modell oktatási folyamatba történő becsatornázására.

5 Tanulmányi és érettségi adatok

5.1 Nyers adatok bemutatása

Tanulmányi adatok

Elemzéseim két adatkör nyers adataira épülnek, az e-naplóból kinyert tanulmányi eredményekre, valamint a manuálisan digitalizált érettségi eredményekre.

Az e-naplóból származó adatok az eredeti adatbázis xlsx formátumú exportálásából jöttek létre, mely folyamat során a különböző tanévek egyes osztályainak adatai szeparált adattáblába kerültek. A tanulmányi adattáblákat jellemezően rendezetlenek voltak, az exportált eredményben véletlenszerűen rendeződtek az egyes tantárgyak, valamint az adott osztály osztályfőnöke döntésének megfelelően tartalmaztak kiegészítő tárgyakat (fakultációk és szakkörök), valamint magatartás és szorgalomjegyeket.

A tanulmányi eredmények további sajátossága, hogy az intézmény minden évfolyamának különbözik a tanterve, vagyis nincsen két olyan évfolyam, amelyben a diákok pontosan ugyanazokat a tárgyakat tanulnák. Ebből következően az adattáblák rendezésekor figyelembe kellett venni az egységesítéshez szükséges kritériumok között a nem tanult tárgyak által okozott nehézséget is.

A tanulmányi eredményeket a 2012-2017. közötti időszakra összesen 144 adattábla rögzítette.

Érettségi adatok

Az érettségi eredményekkel kapcsolatos elsődleges nehézséget az jelentette, hogy az adatok papír alapúból digitálissá történő konverzióját manuálisan kellett elvégezni, mert erről nem állt korábban rendelkezésre digitális példány, az adatok szenzitivitása miatt pedig a tárolási helyiséget digitalizálás céljából sem lehetett elhagyni. Munkám eredményeként egyetlen adattáblában összesítve kerültek az adatok, melyek a feldolgozás folyamatából kifolyólag mind szintaktikai, mind szemantikai hibát is tartalmazhattak.

Az érettségi adatok sajátossága, hogy az egyes diákok különböző számú, szintű (középszint, emelt szint) és kategóriájú (előrehozott, szintemelő, javító) vizsgát tehettek, az előírt kötelező

érettségi tárgyak mellett. Az alábbi táblázat egy rövid statisztikát tartalmaz az érettségi adatok sokszínűségéről (4. táblázat).

Nyers érettségi adatok
• 2011-2016. között összesen 362 diák tett érettségit • Ez alatt az idő alatt 1741 érettségi eredmény született • Átlagosan egy diák 4,81 db érettségi vizsgát tett • A legkevesebb egy diákhhoz tartozó érettségi 1 db, a legtöbb vizsga 7 db volt a vizsgált időszakban • Az 1741 érettségi vizsga szint szerinti megoszlása: 1172 középszintű, 569 emelt szintű érettségi • 16 különböző érettségi tárgy

4. táblázat: Statisztika a nyers érettségi adatokról (Forrás: saját szerkesztés)

Az érettségi eredmények egy táblában szerepelve, a 2011-2016. évekre vonatkozóan foglalják össze a diákok érettségi adatait.

Az érettségi eredményeket 2011-2016 közötti időszakra egyetlen tábla 1741 rekordja rögzítette.

5.2 Az elemzéshez szükséges adatforma

Az adatköröket vizsgálva megállapíthatjuk, hogy a nyers adatok rengeteg olyan attribútumot tartalmaznak, amely nem tartozik szorosan az elemzési témához (pl. szülők neve, lakhely, érettségi éve, előrehozott-e az érettségi, stb.). Első körben tehát ezeket a modell felállításának szempontjából felesleges adatokat kell leválasztanunk a központi résztől.

Fontos azt is megvizsgálnunk, hogy a rendelkezésre álló adatok alkalmasak-e adatelemzésre, vagy azokat át kell ehhez alakítani (pl. külön oszlopban szereplő tárgynevek és hozzá tartozó érdemjegyek egy oszlopba történő összevonása).

Ezután azt kell mérlegelni, hogy melyek azok az attribútumok, amelyek az elemzés szempontjából nem relevánsak, vagyis attribútum aggregálást kell végrehajtani. E folyamatban lehetőség nyílik az egyes attribútumok összevonására (pl. matematika, fizika, földrajz, kémia és biológia tantárgyak eredményeit egységesen reál tárgyakból elért eredményként ábrázolni), elhagyására (pl. előrehozott érettségi, amely időbelisége érdemben nem befolyásolja az eredményességet) vagy normalizálására (pl. hiányzások abszolút értékének levetítése az adott tanuló átlagos éveire). Itt fontos megemlíteni, hogy az elemzés szempontjából szükség van az

egyes eredmények objektív összehasonlíthatóságára. Mivel mind a különböző tárgyak, mind az egyes vizsgaszintek különböző elvárásokat támasztanak a diákok felé (mind tanulmányi, mind érettségi szempontból), ezért nem lehet őket egy az egyben megfeleltetni egymásnak. A dolgozatban azt az általános elfogadott módszert használtam, amit a felvételi során is alkalmaznak a pontok meghatározására, ezt egy új attribútum feltüntetésével (érettségi pontok) ültettem át a gyakorlatba. Ezen érték kiszámítása a 2018-ban hatályos érettségi pont meghatározásával egyezik meg (Felvi.hu, 2018).

Végül azt is meg kell vizsgálnunk, mi az, ami mentén illeszteni tudjuk az adatainkat. Erre a legalkalmasabbak a diákok lesznek, ők szerepelnek mind a két adatbázisban, és hozzájuk társítjuk a két adatkör elemeit. Ez azt jelenti, hogy egységesíteni kell a diákok neveit (pl. harmadik nevek, elgépelésből származó hibák), valamint meg kell határozni, mi történjen azokkal a diákokkal, akikről csak az egyik adatkörben rendelkezünk adattal. A dolgozat során ez utóbbi típusú adatokat végül figyelmen kívül hagytam. Fontos kérdés volt az intézménytől, hogy az adatelemzési fázistól kezdve anonim adatokkal dolgozzak, így végül ezt is meg kell tennünk.

A szükséges adatforma eléréséhez tehát az alábbi lépésekre van szükség:

- 1) a két adatkör irreleváns adatainak eltávolítása
- 2) az egyes attribútumok elemzésre alkalmassá tétele
- 3) az egyes attribútumok becsült faktoranalízise, attribútumok véglegesítése
- 4) a két adatkör illesztésének lehetővé tétele egységesítéssel és adattörléssel
- 5) az adatkörök illesztése, az adatok anonimizálása

5.3 Az adatok elemzésre vetített korlátai

Az adatokból egyértelműen kitűnnek bizonyos elemzési korlátok.

Elsőként a minta nem reprezentatív volta. Mivel az adatokat egyetlen intézmény szolgáltatta, melynek ráadásul speciális helyzete van az oktatási rendszerben (nem állami fenntartású, nyolc osztályos gimnázium), így az elemzésből levont következtetések nem tekinthetők általános érvényűnek.

Másodsorban, a minta kis elemszáma (~300 fő) nem teszi lehetővé a messzemenő következtetések levonását. Ezt fokozza azon adatok eltűnése, amelyek nem felelnek meg az illesztési követelményeknek.

Végezetül azt is figyelembe kell venni, hogy az egyes tanulmányi adatok lépcsőzetes időbeni eltolódása, valamint a nyolc évfolyamos tanulmányi idő miatt nem állnak rendelkezésre azonos időtartamra vonatkozó évfolyam adatok, hiszen az e-napló csupán hat éve működik az intézményben. Az ebből fakadó nehézségekre a dolgozat későbbi részében még részletesen kitérek.

Mindezen korlátok azonban csak azt jelentik, hogy a felállított modell értékelésekor ezeket a szempontokat figyelembe kell venni, a modell által nyújtott eredményeket a megfelelő kontextusban kell értelmezni. A korlátok miatt a végleges eredményeket tehát nem tudományos mérésként, hanem az osztályzat elemzés megvalósítási koncepciójaként kell tekinteni.

6 Az adatelőkészítés folyamata

6.1 Irreleváns adatok eltávolítása

Elsőként a tanulmányi eredmények adattábláiból kell kiszűrni a nem releváns adatokat. A tanulmányi eredményeket tartalmazó nyers adattáblák az alábbi attribútumokat rögzítették:

- belső azonosítószám
- személyes adatok (családi- és utónév, születési hely és idő, lakhely, szülők családi- és utóneve)
- osztályjel
- igazoltan és igazolatlanul távol töltött órák száma
- tantárgyak nevei és megszerzett érdemjegy szövegesen
- magatartás és szorgalom érdemjegy szövegesen
- fakultációk és egyéb szakkörök eredményei (opcionális)

Ezen adatokból az adatbányászathoz az alábbi szűkített listát hoztam létre:

- tanuló neve, születési ideje
- igazoltan és igazolatlanul távol töltött órák száma
- tantárgyakból szerzett érdemjegyek szövegesen
- magatartás és szorgalom érdemjegyek szövegesen

Az érettségi eredményeket rögzítő nyers adattáblák a következő attribútumokat tartalmazták:

- érettségi éve
- vizsgát tevő neve
- tantárgy neve
- előrehozott-e a vizsga
- emelt szintű-e a vizsga
- elért százalékos eredmény
- elért érdemjegy

Ezen adatokból a következő szűk lista keletkezett:

- tanuló neve

- tanuló érettségi pontszáma

6.2 Egyes attribútumok elemzésre alkalmassá tétele

A tanulmányi eredmények adatsoránál a legfontosabb átalakítást a tantárgyaknál kellett elvégezni. A kiinduló állapot (4. ábra) külön oszlopban tartotta számon a tantárgyakat és a szerzett érdemjegyet, valamint az egyes tárgyak sorrendje is véletlenszerű volt.

Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
targy1_jeg	targy2_nei	targy2_or	targy2_jeg	targy3_nei	targy3_or	targy3_jeg	targy4_nei	targy4_or	targy4_jeg	targy5_nei	targy5_or	targy5_jeg	targy6_nei	targy6_or	targy6_jeg
közepes	magyar iro	0	közepes	történelen	0	elégséges	angol nyelv	0	jó	latin nyelv	0	elégséges	matematika	0	elégséges
közepes	magyar iro	0	közepes	történelen	0	közepes	angol nyelv	0	elégséges	latin nyelv	0	elégséges	matematika	0	elégséges
jeles	magyar iro	0	jeles	történelen	0	jó	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	közepes
közepes	magyar iro	0	jó	történelen	0	jó	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	jó
közepes	magyar iro	0	közepes	történelen	0	jó	angol nyelv	0	közepes	latin nyelv	0	közepes	matematika	0	közepes
jeles	magyar iro	0	jeles	történelen	0	jeles	angol nyelv	0	jó	latin nyelv	0	jeles	matematika	0	jeles
jó	magyar iro	0	jó	történelen	0	jó	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	jeles
közepes	magyar iro	0	jeles	történelen	0	közepes	angol nyelv	0	közepes	latin nyelv	0	közepes	matematika	0	elégséges
közepes	magyar iro	0	jó	történelen	0	elégséges	angol nyelv	0	közepes	latin nyelv	0	közepes	matematika	0	közepes
jó	magyar iro	0	jó	történelen	0	jeles	angol nyelv	0	jó	latin nyelv	0	jeles	matematika	0	közepes
jó	magyar iro	0	jó	történelen	0	közepes	angol nyelv	0	jó	latin nyelv	0	jó	matematika	0	jó
jó	magyar iro	0	jó	történelen	0	közepes	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	jó
jó	magyar iro	0	jó	történelen	0	közepes	angol nyelv	0	jeles	latin nyelv	0	közepes	matematika	0	közepes
jeles	magyar iro	0	jeles	történelen	0	jó	angol nyelv	0	jó	latin nyelv	0	jeles	matematika	0	jó
jó	magyar iro	0	jó	történelen	0	jeles	angol nyelv	0	jó	latin nyelv	0	jó	matematika	0	jeles
jó	magyar iro	0	jeles	történelen	0	közepes	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	jó
jó	magyar iro	0	jó	történelen	0	jó	angol nyelv	0	jó	latin nyelv	0	jó	matematika	0	jó
jeles	magyar iro	0	jó	történelen	0	jó	angol nyelv	0	jeles	latin nyelv	0	felmentett	matematika	0	jeles
jeles	magyar iro	0	jó	történelen	0	jó	angol nyelv	0	jó	latin nyelv	0	jó	matematika	0	jó
jó	magyar iro	0	jó	történelen	0	jeles	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	közepes
közepes	magyar iro	0	jó	történelen	0	közepes	angol nyelv	0	közepes	latin nyelv	0	közepes	matematika	0	közepes
elégséges	magyar iro	0	jó	történelen	0	közepes	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	közepes
közepes	magyar iro	0	jó	történelen	0	közepes	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	jó
jó	magyar iro	0	jó	történelen	0	elégséges	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	közepes
elégséges	magyar iro	0	közepes	történelen	0	elégséges	angol nyelv	0	közepes	latin nyelv	0	közepes	matematika	0	közepes
közepes	magyar iro	0	jó	történelen	0	elégséges	angol nyelv	0	jó	latin nyelv	0	közepes	matematika	0	jó
közepes	magyar iro	0	jeles	történelen	0	elégséges	angol nyelv	0	jó	latin nyelv	0	felmentett	matematika	0	közepes

3. ábra: Tanulmányi eredmények egy nyers adattáblája (Forrás: saját szerkesztés)

Az elemezhetőséghez egységesített, minden évfolyamra és osztályra vonatkoztatható adatstruktúra kell, így a következő lépéseket kellett végrehajtani mind a 144 táblán:

- 1) a tantárgyak nevei fejlécként, a szerzett érdemjegy attribútumként szerepeljen az egyes tanulóknál
- 2) a tantárgyakat meghatározott, egységes sorrendben kell feltüntetni (pl. abc-sorrend)
- 3) a tantárgylistát úgy kell meghatározni, hogy minden évfolyamra kitölthető legyen, vagyis az összes tantervben szereplő tárggyal ki kell bővíteni.
- 4) a 144 különböző adattáblát egy közös adattáblában kell rögzíteni.

A folyamat elkészültével egy sztenderdizált, elemezhető adattáblát kapunk (5. ábra).

I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
2012 1 eredmények														
biológia	ének-zene	filozófia	fizika	földrajz	hitman	informatika	kémia	latin nyelv	magatartás	magyar irodalom	magyar nyelv	matematika	műveltség	német nyelv
0	jeles	0	0	0	jeles	0	0	jó	példás	jeles	jeles	jó	0	jeles
0	jeles	0	0	0	jeles	0	0	jó	példás	jeles	jó	jeles	0	jeles
0	jeles	0	0	0	jeles	0	0	jó	változó	jó	jó	jó	0	jó
0	jeles	0	0	0	jeles	0	0	jeles	példás	jeles	jó	elégseges	0	jeles
0	jeles	0	0	0	jeles	0	0	közepes	példás	jó	közepes	közepes	0	jó
0	jeles	0	0	0	jeles	0	0	jó	példás	jeles	jeles	jó	0	jeles
0	jeles	0	0	0	jeles	0	0	jeles	példás	jeles	jeles	jó	0	jó
0	közepes	0	0	0	közepes	0	0	közepes	jó	elégseges	közepes	jó	0	jeles
0	jó	0	0	0	közepes	0	0	jó	példás	jó	jó	jó	0	jó
0	jeles	0	0	0	jó	0	0	jó	példás	jó	jeles	jó	0	jeles
0	jeles	0	0	0	jeles	0	0	jeles	jó	jó	jó	jó	0	jeles
0	jeles	0	0	0	jó	0	0	jó	példás	jó	jeles	jó	0	jó
0	jeles	0	0	0	közepes	0	0	elégseges	példás	közepes	közepes	elégseges	0	jó
0	jó	0	0	0	jó	0	0	jó	példás	jeles	jeles	jó	0	jó
0	jeles	0	0	0	jeles	0	0	jó	példás	jó	jó	jó	0	jeles
0	jó	0	0	0	közepes	0	0	jó	jó	jó	közepes	jó	0	jó
0	jeles	0	0	0	jeles	0	0	jó	változó	jeles	jeles	közepes	0	jeles
0	jeles	0	0	0	közepes	0	0	jó	jó	jeles	jeles	jeles	0	jeles
0	jeles	0	0	0	közepes	0	0	jó	példás	közepes	jeles	jeles	0	jó
0	jó	0	0	0	jó	0	0	jó	példás	jó	jó	jeles	0	jeles
0	jeles	0	0	0	jó	0	0	jó	jó	jeles	jó	jó	0	jeles
0	jeles	0	0	0	jó	0	0	jó	példás	jeles	jó	jó	0	jeles
0	jeles	0	0	0	jeles	0	0	jeles	példás	jó	jeles	jó	0	jeles
0	jeles	0	0	0	jó	0	0	közepes	példás	jó	jó	jó	0	jeles
0	jeles	0	0	0	közepes	0	0	jó	példás	jó	jó	jó	0	jeles
0	jeles	0	0	0	jeles	0	0	jó	példás	jeles	jeles	jó	0	jeles
0	jeles	0	0	0	jó	0	0	jeles	jó	jeles	jeles	jó	0	jeles
0	jeles	0	0	0	jeles	0	0	jeles	példás	jó	jeles	jó	0	jeles
0	jeles	0	0	0	jó	0	0	jó	példás	jó	jó	jó	0	jeles
0	jó	0	0	0	közepes	0	0	közepes	jó	közepes	közepes	jó	0	jeles
0	jeles	0	0	0	jeles	0	0	jeles	példás	jeles	jeles	jó	0	jeles
0	jeles	0	0	0	jeles	0	0	jeles	jó	jeles	jeles	jó	0	jeles
0	jó	0	0	0	jó	0	0	jó	jó	jó	közepes	jó	0	jeles
0	jó	0	0	0	közepes	0	0	közepes	példás	jó	közepes	elégseges	0	jó

4. ábra: Sztenderdizált adattábla (Forrás: saját szerkesztés)

6.3 Az érettségi eredmények objektív összehasonlítása

Az érettségi adatok nyers formátumához képest (6. ábra) az egyetlen átalakítást az egyes érettségi eredmények egyetlen attribútumként történő rögzítése jelentette. Ez azért volt szükséges, hogy a különböző érettségiket objektíven összehasonlíthassuk. Az alapproblémát az jelenti, hogy nincsen rá egységes szabály, milyen viszonyban állnak egymáshoz képest a középszintű és az emelt szintű érettségik. Mivel minden egyes tárgy esetében más-más arányszám fejezné ki a két szint különbségét, és ezek meghatározása nem áll módunkban, ezek egy-egy arányú megfeleltetése nem megoldható. Hogy mégis mérhető legyen az érettségi teljesítmény, a felvételi eljárás során alkalmazott módszert választottam, és minden tanulóra kiszámoltam az érettségi pontot.

A	B	C	D	E	F
Év	Név	Tárgy	Emelt?	Százalék	Érdemjegy
2011		angol nyelv	x	85	5
2011		magyar nyelv és irodalom		72	4
2011		matematika		70	4
2011		német nyelv		91	5
2011		történelem	x	79	5
2011		biológia	x	60	5
2011		magyar nyelv és irodalom		81	5
2011		matematika		74	4
2011		német nyelv		88	5
2011		történelem		89	5
2011		angol nyelv		78	4
2011		informatika	x	89	5
2011		magyar nyelv és irodalom		60	4
2011		matematika		76	4
2011		történelem		56	3
2011		angol nyelv		92	5
2011		magyar nyelv és irodalom		76	4

5. ábra: Nyers érettségi adattábla (Forrás: saját szerkesztés)

A tanuló érettségi pontszámának meghatározása megegyezik a 2018-ban hatályos érettségi pontszám kiszámításával (Felvi.hu, 2018), aminek módja:

$$\begin{aligned}
 &\text{Össz. érettségi pontszám (max. 400 pont)} = \\
 &= \text{Két felvételi tárgy százalékos eredménye (max. 200 pont)} + \\
 &+ \text{Összes érettségi eredmény átlaga (max. 100 pont)} + \\
 &+ \text{Többletpontok az emelt érettségiért (max. 100 pont)}.
 \end{aligned}$$

Ezzel a módszerrel lehetővé vált, hogy azt az optimalizáló szempontot vegyük az elemzés során figyelembe, amit a tanulók tartanak szem előtt az érettségi során, valamint aggregált és egységesített mérőszámot tudunk képezni a különböző vizsgák eredményeiről (7. ábra).

A	B	C	D	E	F	G	H	I
Név	Tárgy	Emelt?	Százalék	Érdemjegy	2 beszámító	Pluszpont össz	Érettségi átlag	Össz pont
	angol nyelv	x	85	5	164	100	79,4	343,4
	történelem	x	79	5				0
	német nyelv		91	5				0
	magyar nyelv és irodalom		72	4				0
	matematika		70	4				0
	biológia	x	60	5	149	50	78,4	277,4
	történelem		89	5				0
	német nyelv		88	5				0
	magyar nyelv és irodalom		81	5				0
	matematika		74	4				0
	informatika	x	89	5	167	50	71,8	288,8
	angol nyelv		78	4				0
	matematika		76	4				0
	magyar nyelv és irodalom		60	4				0
	történelem		56	3				0
	angol nyelv		92	5	182	0	83	265

6. ábra: Érettségi eredmények egységesítése (Forrás: saját szerkesztés)

6.4 Az attribútumok aggregálása

Az összevonás szükségessége

A 3. táblázat adatai alapján is megfigyelhető, hogy az adatok között nem található két évfolyam, melyekről ne különböző mennyiségű adat állna rendelkezésre. Ez egyfelől az intézmény nyolcosztályos jellegéből, másfelől az e-napló hatéves fennállásából következik.

További nehézség, hogy az egyes évfolyamok különböző tanterveket követnek, így például a kilencedik évfolyam nem tanul biológiát, míg tizedik osztály után nincsen a tantervben az informatika tárgya. Ezen összehasonlíthatósági korlátok vezettek ahhoz, hogy az egyes tárgyakat és érdemjegyeiket közvetlen módon nem tudjuk összevetni, így azokat aggregált formában veszem figyelembe a modellépítéskor.

Az összevonás módja

A tanulmányi eredmények adatkörére vonatkozóan, féléves szinten az alábbi aggregálást végeztem (5. táblázat), mely során az egyes attribútumok elemzési relevanciáját felülvizsgáltam, majd a következő kategóriák egyikébe soroltam:

- 1) az attribútum nagyon releváns, változtatás nélkül megmarad (táblázatban: megmarad)
- 2) az attribútum nagyon vagy közepesen releváns, de önmagában nem értelmezhető, átalakításra szorul (táblázatban: átalakítva: „xy”)
- 3) az attribútum nagyon vagy közepesen releváns, de túlságosan részletessé tenné a modellt, így összevonható hasonló attribútumokkal (táblázatban: összevonva itt: „xy”)
- 4) az attribútum kevésbé vagy egyáltalán nem releváns, így törölhető (táblázatban: törölve)

Attribútum neve	Aggregálás eredménye	Indoklás (opcionális)
tanuló neve	megmarad	illesztéshez szükséges
tanuló születési ideje	törölve	a sokévi adatösszevonás miatt irreleváns
tanuló neme	megmarad	
tanuló tárgyévi évfolyama	törölve	a sokévi adatösszevonás miatt irreleváns
tanuló osztálya	megmarad	
igazolt órák száma	átalakítva: „átlagos hiányzás”	aggregált forma
igazolatlan órák száma	törölve	nem hordoz többletinformációt

magatartás érdemjegy	átalakítva: „átlagos magatartás”	aggregált forma
szorgalom érdemjegy	átalakítva: „átlagos szorgalom”	aggregált forma
angol nyelv érdemjegy	összevonva itt: „humán átlag”	a különböző évfolyamok jobb összehasonlíthatóságának érdekében összevont formában szerepelnek
biológia érdemjegy	összevonva itt: „reál átlag”	
éneke-zene érdemjegy	összevonva itt: „képesség átlag”	
filozófia érdemjegy	összevonva itt: „humán átlag”	
fizika érdemjegy	összevonva itt: „reál átlag”	
földrajz érdemjegy	összevonva itt: „reál átlag”	
hittan érdemjegy	összevonva itt: „humán átlag”	
informatika érdemjegy	összevonva itt: „reál átlag”	
kémia érdemjegy	összevonva itt: „reál átlag”	
latin nyelv érdemjegy	összevonva itt: „humán átlag”	
magyar irodalom érdemjegy	összevonva itt: „humán átlag”	
magyar nyelv érdemjegy	összevonva itt: „humán átlag”	
matematika érdemjegy	összevonva itt: „reál átlag”	
művészettörténet érdemjegy	összevonva itt: „humán átlag”	
német nyelv érdemjegy	összevonva itt: „humán átlag”	
rajz érdemjegy	összevonva itt: „képesség átlag”	
természetismeret érdemjegy	összevonva itt: „reál átlag”	
testnevelés érdemjegy	összevonva itt: „képesség átlag”	
történelem érdemjegy	összevonva itt: „humán átlag”	

5. táblázat: Eredeti attribútumok aggregálása (Forrás: saját szerkesztés)

A táblázatban szereplő attribútumok rövid magyarázata:

- *átlagos hiányzás*: a tanuló összes, az adatbázisban rögzített félévét figyelembe véve az egy félévre jutó átlagos hiányzásának kategorizált, numerikus értéke. (0: 60 vagy kevesebb hiányzás, 1: 60 és 130 közötti hiányzás, 2: 130 és 170 közötti hiányzás, 3: 170 és 230 közötti hiányzás, 4: 230 feletti hiányzás. A kategóriák az egész sokaságra vetített eloszlás természetes kvintilisei által meghatározott, kerekített értékei mentén lettek meghatározva.)
- *átlagos magatartás*: a tanuló összes, adatbázisban rögzített félévét figyelembe véve az összes magatartás érdemjegyének átlagolt értéke.
- *átlagos szorgalom*: a tanuló összes, adatbázisban rögzített félévét figyelembe véve az összes szorgalom érdemjegyének átlagolt értéke.
- *humán átlag*: az egyes humán tárgyak (angol nyelv, filozófia, hittan, latin nyelv, magyar nyelv, magyar irodalom, művészettörténet, német nyelv, történelem) érdemjegyeinek átlagolt értéke.
- *reál átlag*: az egyes reál tárgyak (biológia, fizika, földrajz, kémia, informatika, matematika, természetismeret) érdemjegyeinek átlagolt értéke

- *készség átlag*: az egyes készség tárgyak (ének-zene, rajz, testnevelés) átlagolt értéke.

6.5 Az adatkörök illesztési feltételeinek megteremtése

Ahhoz hogy az adattáblákat illeszteni tudjuk, a közös kulcsként választott „tanuló neve” attribútum felülvizsgálata szükséges. Az adatok kinyerésekor manuálisan felvitt rekordoknál potenciális hibák találhatók, például eglépelés, kimaradó harmadik nevek. Ezen felül foglalkozni kell az adatkörök aszimmetrikus voltával, vagyis azzal, hogy bizonyos tanulók csak az egyik rendszerben szerepelnek. Az elemzés szempontjából az előbbire a manuális korrekció, utóbbira az adott tanulók illesztésbeli figyelmen kívül hagyása volt a megoldás.

6.6 Az adatkörök illesztése, adatok anonimizálása

A tanulmányi eredmények és az érettségi eredmények tisztított adattábláját a tanulók neve mentén illesztve kaptuk az elemzés alapjául szolgáló végleges táblát (8. ábra).

Azonosító	nem	osztály	átlagos_hiányzás	átlagos_magatartás	átlagos_szorgalom	reál_átlag	humán_átlag	készség_átlag	érettségi_pont
10175063	2	1	177	5,00	4,50	3,88	4,46	4,63	305
10500809	1	1	65	4,00	3,00	3,08	4,00	4,25	197
10812665	2	1	129	5,00	4,13	3,57	4,31	4,50	241
12198277	1	1	43	3,63	3,75	3,82	4,41	4,27	270
12316619	2	2	129	4,25	4,25	3,96	4,68	4,83	272
12373523	1	1	71	4,63	4,25	3,91	4,72	4,82	318
12714217	2	2	59	5,00	5,00	4,52	4,87	4,91	286
12774181	2	2	137	4,75	4,50	4,04	4,63	4,73	294
13078853	2	2	63	5,00	4,00	3,67	4,13	4,75	234
14694905	1	2	96	2,00	3,50	3,17	3,88	4,50	290
14818425	2	2	146	4,00	3,25	2,79	3,68	3,92	152
15409080	1	1	76	4,33	4,83	4,00	4,59	4,56	304
16048505	2	1	107	5,00	5,00	4,67	4,88	5,00	331
16266912	1	1	92	4,00	3,50	4,00	3,38	4,00	273
16511293	2	1	148	5,00	4,50	3,25	4,70	4,25	293
16704355	1	2	167	4,00	3,00	3,07	3,71	4,14	260
17512991	1	1	123	4,00	4,00	3,67	4,50	4,25	319
17952610	2	2	137	5,00	4,50	3,83	4,75	5,00	312
18053720	2	1	105	5,00	5,00	4,67	4,94	5,00	244
18355377	2	2	89	5,00	5,00	4,67	4,67	4,88	297
18374627	1	1	47	4,00	4,00	4,22	4,09	4,75	261
19095461	2	1	77	4,50	4,25	3,65	4,53	5,00	235
19221219	2	1	97	5,00	5,00	4,33	4,80	4,75	299
19325296	1	1	223	4,50	3,00	3,17	4,13	4,50	259

7. ábra: Véglegesített, illesztett adattábla (Forrás: saját szerkesztés)

Az adatok előkészítésének utolsó lépéseként az adatokat anonimizáltam, melyhez egy segédtáblát alkalmaztam, ebben szerepeltek a véletlen kulcsokként generált egyedi azonosítók.

A fenti lépések elvégzésével létrejött az az egységesített, normalizált és tisztított adatállomány, amely alkalmas az adatbányászati vizsgálatok elvégzésére.

6.7 Egy adatelőkészítési javaslat a jövőbeli elemzésekhez

Az adattisztítás és adatelőkészítés folyamatának nehézségei két okra vezethetők vissza. Először is, nem volt egységes az e-naplóból kinyert adatok formátuma, ami az exportálás szűk beállítási lehetőségeinek köszönhető. Erre, az e-napló szabad és nyílt forráskódú szoftver lévén, megoldás lehet egy olyan bővítmény fejlesztése, ami az exportálás formátumát engedi szerkeszteni.

Másodszor, az érettségi adatoknak nem létezett digitális formátuma. Erre az a megoldás, hogy az adatok rendelkezésre állásának évében (vizsga tárgyéve) a digitális adatbázist rögzíteni kell. Visszamenőlegesen pedig felhasználható a dolgozathoz előállított adattábla.

7 Osztályzáselemzés prediktív módszerrel

7.1 Az alkalmazott modell típusok bemutatása

Ahogy azt a bevezetőben is tárgyaltuk, Baker és Siemens (2014) az előrejelző modelleknek három fő típusát különbözteti meg. Ezek az osztályozás, a regresszió és a rejtett tudás becslés. A prediktív módszer célja mindig egy adattényező (célváltozó) értékére becslést adni az adathalmaz többi részének bizonyos kombinációit elemezve.

Az *osztályozó* modellnél a célváltozó lehet bináris (pl. 0 vagy 1), vagy kategóriai változó. A leggyakoribb eszközei közé tartozik a döntési fa vagy a véletlen erdő. A *regresszió*nál a célváltozó egy folytonos érték, például egy szám. A leggyakoribb a lineáris regresszió. A harmadik típus a *rejtett tudás becslés*. Ez a tanulók tudását vagy speciális képességeit és koncepcióit veti össze a kritikus képességek helyességével.

A dolgozat kérdésének fényében beláthatjuk, hogy a folytonos értékként megjelenő érettségi pontok előrejelzését keresve, az osztályozó és a rejtett tudást becslő modellek alkalmazásához másféle bemeneti vagy kimeneti feltételek kelljenek. Így az alkalmazott prediktív modell esetünkben a regresszió lesz.

A regressziós modell esetében többféle megközelítés is alkalmazható, a dolgozathoz használt adatbányászati szoftver, a RapidMiner (2018) többek között az alábbi lehetőségeket kínálja regressziós elemzésre:

- 1) GLM (Generalized Linear Model, vagyis Általános Lineáris Modell)
- 2) Lineáris regresszió
- 3) Deep Learning

GLM: a tradicionális lineáris modell egy kiterjesztése. Az algoritmus különböző általános lineáris modelleket illeszt az adathalmazra a logaritmikus hasonlóság maximumát keresve. Az illeszkedés-keresés párhuzamosan folyik.

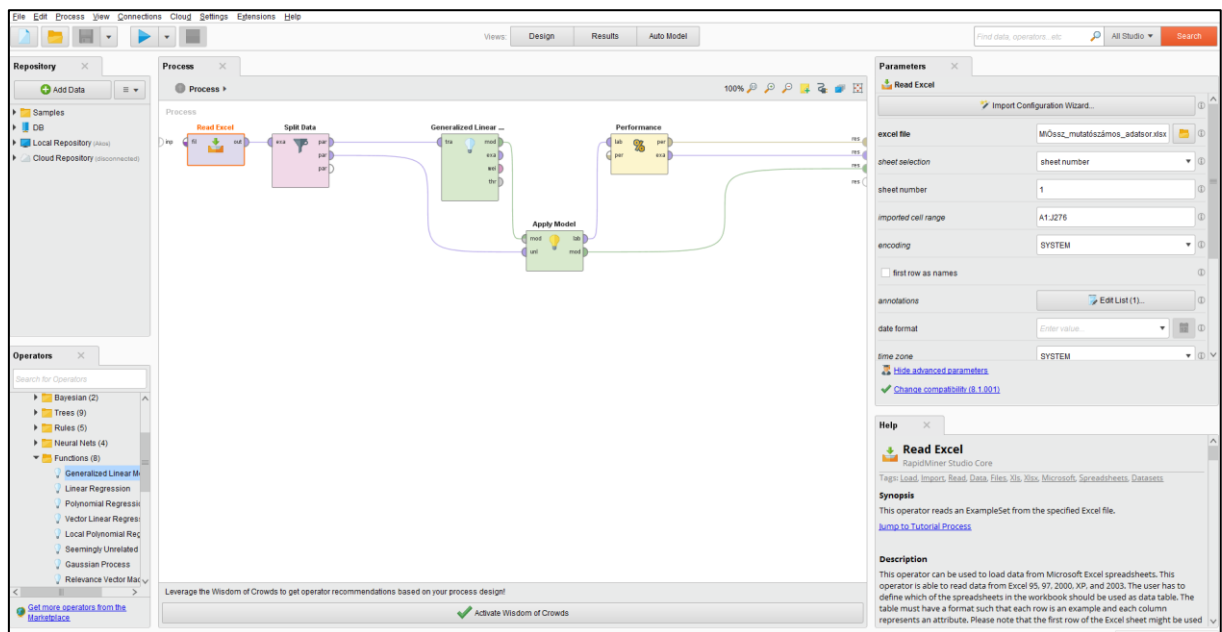
Lineáris regresszió: lineáris egyenletet próbál létrehozni a skaláris célváltozó és a különböző magyarázóváltozók között. A modellválasztáshoz az Akaike-kritériumot alkalmazza, amely egy relatív illeszkedést vizsgáló módszer.

Deep Learning: a RapidMiner többretegű feed-forward mesterséges neuronhálója, ami a H2O platformon alapuló algoritmust használja.

7.2 Modellezési folyamat

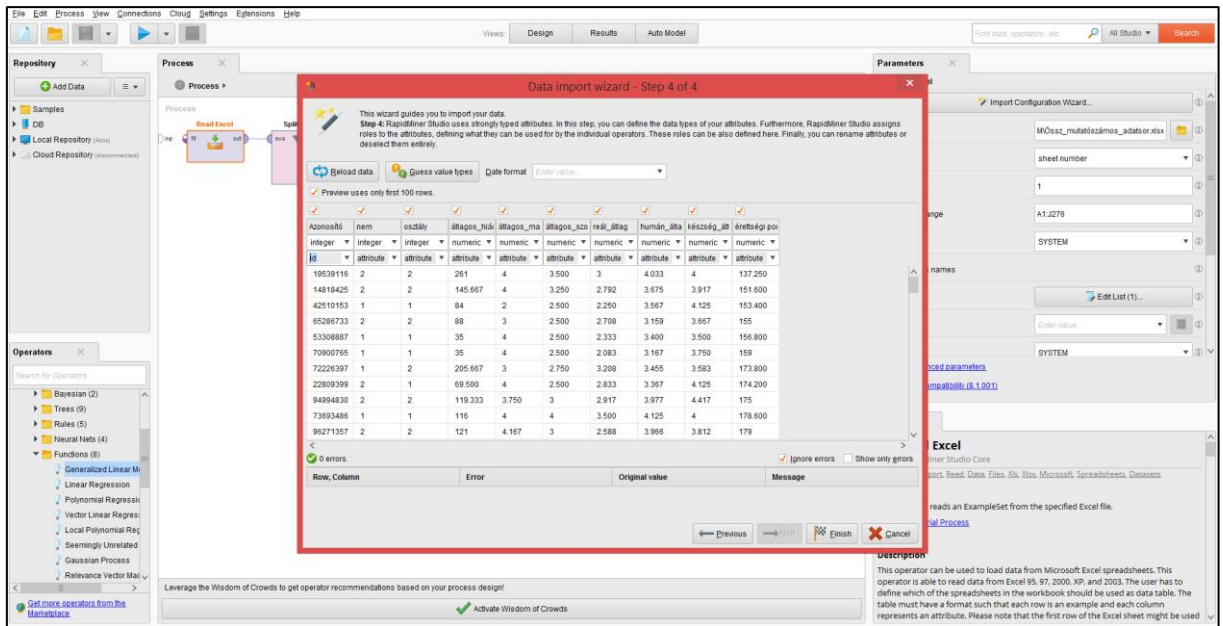
Az egyes modellek elkészítésének folyamata az elemző szoftverben három lépcsőben történt.

Elsőként a modell struktúráját és az alapbeállításokat kellett meghatározni (8. ábra)



8. ábra: RapidMiner modell struktúra létrehozása (Forrás: saját szerkesztés)

Második lépésként az adatokat kellett bevinni a modellbe, valamint az adattípusokat kellett beállítani (9. ábra).



9. ábra: RapidMiner modell adatok felvétele (Forrás: saját szerkesztés)

Végül a kapott eredmény értelmezése következett (10. ábra).

Row No.	Azonosító	érettségi pont	prediction(átlag)	nem	osztály	átlagos_hő	átlagos_ma	átlagos_szo	reál_átlag	humán_átlag	készség_átlag
1	65289733	155	210.963	2	2	88	3	2.500	2.708	3.159	3.667
2	94994830	175	232.756	2	2	119.333	3.750	3	2.917	3.977	4.417
3	80072308	204	256.090	1	1	145.750	4.500	3.667	3.294	4	4.438
4	33263194	205.200	258.035	2	1	224	4.333	4.167	3.265	4.130	4.375
5	73061883	205.200	238.459	2	1	96.750	4	3.833	3.029	3.827	4.875
6	64489142	207.600	256.342	1	1	97.567	4	3.250	3.478	3.967	4.333
7	84521382	208.800	257.721	1	2	81	4.250	3.500	3.283	4.243	4.045
8	91346071	212	246.184	2	2	103	5	4.500	2.833	4.438	4.500
9	41405554	224	214.891	1	2	83	2	2.500	2.333	3.525	4.500
10	82638598	224.400	233.201	2	2	199.500	4.167	3.667	2.794	3.793	4
11	65005172	233.800	256.557	2	2	91.400	4.375	3.875	3.410	4.286	4.409
12	70095654	235.167	263.112	2	1	107	5	4	3.417	4.571	4.750
13	10812655	241	265.202	2	1	129.200	5	4.125	3.585	4.314	4.500
14	52047098	242.800	256.153	2	2	204.500	4	3.500	3.324	4.276	4.438
15	80384700	243.900	236.210	1	1	6	4	3	3.167	3.438	3.750
16	72479106	246.200	274.427	2	1	85	4.500	4	4	4.168	4.750
17	30550226	256	279.185	1	2	90	4	4.500	3.667	4.375	4.500
18	77969431	262.200	242.973	2	2	48	4.500	4	2.833	4.438	4.250
19	30031476	263.200	263.163	1	1	119	4	3.500	3.500	4.067	4.375
20	65828234	264	262.079	1	1	97.600	3.750	3.875	3.783	4.471	4.455
21	94285583	266	235.173	2	2	62	5	3	3.333	3.688	4.250
22	38359114	271.600	271.263	1	1	105.500	4	3.833	3.559	4.212	3.938
23	66579514	272.400	317.241	2	1	16	5	5	4.833	4.938	5
24	47221307	273	285.486	1	1	108.333	4.500	4	4	4.444	5
25	16269912	273.400	265.791	1	1	92	4	3.500	4	3.375	4
26	83281903	278.600	258.948	2	1	162	5	4	3.333	4.400	4.750
27	26309114	280	309.636	1	1	78.667	4.500	5	4.667	5	4.667

10. ábra: RapidMiner modell eredmények (Forrás: saját szerkesztés)

Az alábbi logikát követve végeztem el a különböző méréseket az egyes modell típusokra.

7.3 Választás a prediktív modellek között

A regressziós modellek értékelésére több módszer is létezik, többek között az Osten (1988) által ajánlott korrelációs együttható, és a modell hibájának vizsgálata. A dolgozatban az alábbi mutatókat vizsgálom:

- 1) *Determinációs együttható (R^2 mutató):* százalékosan értelmezhető mutatószám. Kifejezi, hogy a magyarázó változók együttesen milyen hányadban (mekkora százalékban) magyarázzák a célváltozó varianciáját.
- 2) *Átlagos négyzetes hiba (MSE mutató):* mutatószám, amely a modell varianciáját és a torzítását együttesen veszi figyelembe, ezáltal objektíven összemérhetővé teszi a különböző modelleket.

Az alábbi táblázatban (6. táblázat) a három prediktív modell mutatószámait hasonlítom össze különböző adatfelosztási arányok mellett. Az adatfelosztási arány az a mutatószám, amely meghatározza a tanító és a tesztelő adathalmaz arányát a rendelkezésre álló sokaságra vetítve.

Mutatószám / modell	GLM	Lineáris regresszió	Deep Learning
<i>Adatfelosztási arány: 0,6 – 0,4.</i>			
R² (%)	49,9%	42,5%	31,2%
MSE	1146,884	1562,542	1962,844
<i>Adatfelosztási arány: 0,67 – 0,33</i>			
R² (%)	47,7%	45,3%	42,0%
MSE	1238,668	1464,593	2886,268
<i>Adatfelosztási arány: 0,75 – 0,25.</i>			
R² (%)	47,4%	47,6%	34,3%
MSE	1248,393	1425,591	2959,360
<i>Adatfelosztási arány: 0,8 – 0,2.</i>			
R² (%)	48,2%	46,6%	35,5%
MSE	1218,657	1570,537	3783,357

6. táblázat: Prediktív modellek összehasonlítása (Forrás: saját szerkesztés)

A 6. táblázat eredményeiből a következőket olvashatjuk le:

- 1) A Deep Learning minden tekintetben elmaradt a másik két modelltől, így alkalmazását elvethetjük.

- 2) A négy mérésből három esetben a GLM módszer mutatta a legmagasabb eredményt, egy esetben a Lineáris regresszió. Ugyanakkor az MSE mutatót tekintve a Lineáris regresszió mindegyik esetben nagyobb értékeket vett fel, mint a GLM, így alkalmazását elvethetjük.

Ezen megfigyelések alapján kijelenthetjük, hogy összességében a GLM módszer mutatja a legnagyobb magyarázó erőt a modell és a célváltozó között. A legmagasabb értéket ugyan nem mindig ebben a modellben kaptuk, de összességében stabilan magasabb összefüggést tudott kimutatni, alacsonyabb MSE mutató mellett.

A prediktív modellek közül ezért a GLM módszert alkalmazom a dolgozatban.

7.4 Modell felépítése, kapott eredmény bemutatása

A 6. táblázatban megfigyelhettük, hogy az adatfelosztás arányától nagyban függ a modell magyarázó ereje.

Ennek feloldására olyan validálási módszert kell keresni, amely kevésbé érzékeny a manuális beállításokra. Kohavi (1995) munkájában leírja, hogy kisméretű adathalmazok esetében a manuális validálás helyett sokszor a kereszt-validációt alkalmazzák, mely több, párhuzamos validáció mentén segít stabilabbá tenni a modell értékelését. Éppen ezért a kiválasztott GLM modellen a gyakorlatban elterjedt 10 körös kereszt-validációt is alkalmaztam (7. táblázat).

Mutatószám / modell	GLM modell
<i>10 körös kereszt-validáció</i>	
R² (%)	48,39%
MSE	1268,6378

7. táblázat: Kereszt-validálással kapott eredmény (Forrás: saját szerkesztés)

Megfigyelhetjük, hogy a 10 körös kereszt-validáció során a végső determinációs együttható értéke elmarad a GLM modell manuális beállításoknál mért legmagasabb értékétől, mégis a legtöbb mért értéknél magasabb.

Ennek ismeretében az ehhez az értékhez tartozó modellt választhatjuk, melynek együtthatóit a 8. táblázat mutatja be.

Attribútum	Együttható értéke
nem	-13,261
osztály	1,281
átlagos hiányzás	0,044
átlagos magatartás	0,491
átlagos szorgalom	7,040
reál átlag	30,205
humán átlag	25,510
készség átlag	-10,184
tengelymetszet	79,608

8. táblázat: A GLM modell együtthatói (Forrás: saját szerkesztés)

Az egyes együtthatók értékét úgy értelmezzük, hogy az adott attribútum egységnyi változása mekkora változást jelentene a célváltozóban, a többi attribútum értékének állandósága mellett, vagyis:

- 1) a *nem* (bináris) változó egységnyi növekedése a célváltozó 13,261-kel való csökkenését vonja maga után. Vagyis a modell szerint a többi attribútum változatlansága mellett a fiúk (1-es érték) átlagosan 13,261 ponttal érnek el magasabb eredményt ugyanazon bemeneti feltételek mellett, mint a lányok (2-es érték).
- 2) az *osztály* (bináris) változó egységnyi növekedése a célváltozó 1,281-kel növekedését vonja maga után. Az 1-es értéket az „A” osztályosok, a 2-es értéket a „B” osztályosok jelentették, vagyis más változók változatlansága mellett a „B” osztályosok 1,281 ponttal magasabb érettségi pontszámot szereznek a modell szerint.
- 3) az *átlagos hiányzás* (kategrikus) változó egységnyi növekedése, a fentiekhez hasonló körülmények esetén 0,044 pontnyi növekedést jelent a modell szerint.
- 4) az *átlagos magatartás* (folytonos) változó egységnyi növekedése a fenti feltételek mentén 0,491 pontnyi növekedést jelentene.
- 5) az *átlagos szorgalom* (folytonos) változó egységnyi növekedése a fenti feltételek mentén 7,04 pontnyi növekedést jelentene.
- 6) a *reál átlag* (folytonos) egységnyi növekedése 30,205 pontnyi növekedést okozna a többi változó változatlansága esetén a modell szerint. Érdemes figyelembe venni, hogy a többi attribútumhoz képest milyen nagy súllyal rendelkezik ez a változó.
- 7) a *humán átlag* (folytonos) egységnyi növekedése 25,510-kel növelné a célváltozó értékét a fenti feltételek mentén.

- 8) a *készség átlag* (folytonos) változó egységnyi növekedése 10,184-kel csökkentené a célváltozó értékét, ami megint érdemi információt hordozhat magában.

8 A prediktív modell értékelése

8.1 Mutatószámok szerepe az értékelésben

A kapott modell értékeléséhez fontos megvizsgálunk azt objektív, statisztikai szempontok szerint is. Ehhez olyan mérőszámokat választottam, melyek gyakran alkalmazott mutatók modellértékeléskor, valamint kiszámításukat automatikusan elvégezte az elemző szoftver.

Az egyes mutatószámok meghatározásai Hunyadi és Vita (1991 és 1992) művei alapján kerülnek bemutatásra.

Relatív hiba

A relatív hiba százalékosan értelmezhető mutatószám. A modell abszolút hibájának és a kiinduló adatoknak a hányadosa, a modell esetében tehát az átlagos becsült érettségi pont és az eredeti adatok számított érettségi pont eltérésének, valamint a becsült értékek átlagának a hányadosa.

Akkor mondhatjuk, hogy a modell elfogadható, ha a relatív hiba kisebb a relatív hibakorlátnál. Ez utóbbi meghatározása az abszolút hibakorlátra épül, amely a mérés pontosságától függ. Mivel egész értékekről beszélünk, így a fél pontos eltérés vehető az abszolút hibakorlátnak.

$$\text{Ez alapján: relatív hibakorlát} = \frac{\text{abszolút hibakorlát}}{\text{becsült érték}} = \frac{0,5}{277} = 0,18\%$$

A modellnél mért relatív hiba: 11,0 % +/- 2,27%, ami nagyságrendekkel elmarad az egész számra való pontos mérés megengedett hibakorlátjától.

Ugyanakkor a becslés nem egy mérési folyamat, így ez önmagában még nem jelent semmit. Ha visszafele számolunk, megkaphatjuk, a 11,0%-os mért relatív hiba érvényes hibakorláttal mekkora abszolút hibát eredményezne.

$$\text{abszolút hiba} = \frac{\text{relatív hibakorlát} * \text{becsült érték}}{100} = \frac{11,0\% * 277}{100} = 30,47.$$

Ez azt jelenti, a modell becslése kb. egy 30-as pontossággal vett statisztikai mérésnek felel meg. Ez a 400 maximum értékű érettségi ponthatárhoz képest egy nagyon durva becslést jelent.

Átlagos négyzetes hiba (MSE)

Az átlagos négyzetes hiba a variancia és a torzítás, mint két lehetséges becslési tulajdonság objektív összehasonlításán alapszik. A mutatószám alapvetően különböző becslések összehasonlítására alkalmas, így a modell értékeléséhez választanunk kell egy referencia modellt, amihez viszonyíthatunk. Két modell közül pontosabb becslést ad az, amelyiknek kisebb az átlagos négyzetes hibája.

Négy referenciamodellt állítottam fel az értékeléshez.

- 1) minden eddigi diák érettségi pontját a teljes sokaságra vett átlaggal becsüljük (Referencia 1)
- 2) az érettségi pontot a tanulmányi átlagok összegével egyenesen arányosan becsüljük (Referencia 2)
- 3) az ismert érettségi ponttól vett 20%-os véletlen eltérést alkalmazzuk becsült értéknek (Referencia 3)
- 4) az ismert érettségi pont tízes kerekítését alkalmazzuk becsült értéknek (Referencia 4)

A kapott eredményeket a 9. táblázat foglalja össze.

Modell	MSE
Referencia 1	2458,60
Referencia 2	4975,88
Referencia 3	1170,70
Referencia 4	9,15
GLM modell	1349,97

9. táblázat: MSE érték vizsgálat referencia modellekkel (Forrás: saját szerkesztés)

A táblázatból kitűnik, hogy a modell pontosabb becslést ad, mint az első két referencia modell. Ugyanakkor az is látszik, hogy nagyságrendileg eltér a negyedik referenciaként választott

modelltől. A harmadik referenciamodell MSE értéke áll a legközelebb a GLM mutatójához, vagyis azt tudjuk mondani, hogy nagyságrendileg közel azonos pontossággal becsülhetnek.

Determinációs együttható (R^2)

A determinációs együttható, hasonlóan a relatív hibához, egy százalékosan értelmezhető mutató. Megmutatja, hogy a független változók együttesen milyen erősséggel határozzák meg a célváltozó ingadozását. Másképpen fogalmazva azt fejezi ki, a függő változó teljes varianciája mekkora arányban tudható be a regressziónak, vagyis a modell attribútumainak.

Ez az érték a modell 10 körös kereszt-validációja során kiszámított 48,39%.

Vagyis a felállított modellben a tanulókról gyűjtött adatok a végeredményt 48,39%-ban magyarázzák. A függvényyszerű megfeleltetéshez szükséges maradék 51,61%-ot nem magyarázza a modell.

8.2 A prediktív modell eredményének értékelése

A modell az egyes mutatószámokkal való összevetésben megmutatta, hogy bár a nagyon pontos statisztikai méréseknek nem megfelelő mértékben jelzi előre az érettségi eredményeket, jóval pontosabb a véletlenszerű, vagy nagy vonalakban átlagoló eredményeknél.

Az összevetés eredményeit többek között magyarázzák a dolgozat elején megfogalmazott adatkinyerési és környezeti korlátok, hiszen ezek csökkentették a minta feldolgozási lehetőségeit, és a következtetések levonási lehetőségeit.

Mindezek fényében belátható, hogy a modell nem alkalmas széleskörű alkalmazásra. A kezdeti célra, vagyis az osztályzat elemzési koncepció relevanciájának és lehetőségeinek bemutatására azonban alkalmas az eredmény.

Összefüggés a tanulmányi és érettségi eredmények között

A dolgozat elején az alábbi kérdést tettem fel:

*Milyen összefüggés mutatható ki a tanulmányi és érettségi eredmények között,
és ezt az ismeretet hogyan lehet az oktatásban közvetlenül kamatoztatni?*

A kérdésre válaszképpen egy prediktív modellt állítottam fel, az osztályzat elemzés segítségével. A modell magyarázóereje 48,39%-os lett.

A modellt kiértékeltem különböző mutatószámok segítségével, majd összevettem az oktatási adatbányászat különböző részterületein elért tudományos munkák eredményeivel.

Az adatok származási környezetét tekintve, miszerint a mintavétel nem reprezentatív, valamint figyelembe véve a minta elemszámának méretét és a rendelkezésre álló adatok hiányosságait, meg kell állapítanunk, hogy az eredményt csupán egy koncepció felállításának első lépéseként lehet értelmezni, abból messzemenő következtetéseket levonni nem lehet.

Ugyanakkor a modell egyértelműen kimutatta, hogy az adott intézményben a vizsgált évfolyamoknál megfigyelhető egy közepesen erős, 48,39%-os magyarázóerő a tanulmányi adatok és az érettségi eredmények között. Ez a determinációs együttható kétféle jelentést is hordoz magában. Egyfelől rávilágít arra, hogy a kapcsolat jelen van, mégpedig kimutatható és mérhető formában. Másfelől azt is megerősíti, hogy a kapcsolat nem függvénytípusú, sőt a vizsgált szituációban a magyarázóerő közel felét szolgáltatja csak. Ez azt jelenti, hogy az érettségi eredményeket a tanulmányi adatok mellett még legalább ugyanekkora mértékben sok más tényező is befolyásolja.

8.3 Tanulmányi eredmények szerepének újraértékelése klaszterezéssel

Intuitív előítéletek a tanulmányi eredményekről

A középiskolákban az az általános nézet figyelhető meg mind a tanárok, mind a diákok részéről, hogy a tanulmányi eredmények egyértelműen előre vetítik az érettségi eredményeket. A kiemelkedő, végig kitűnő tanulótól egyértelműen kitűnő érettségit várnak, míg ezzel párhuzamosan a gyengébb osztályzatokat szerző tanulók eredményeit előre lesajnálják.

Ezt a sztereotip nézetet választottam a modell gyakorlatba ültetési lehetőségeinek egyik bemutatásaként. Ha a nézet által megfogalmazott ok-okozati kapcsolatot lefordítjuk az adatbányászat nyelvére, akkor azt mondhatjuk, hogy a modell magyarázóerejének növekednie kell abban az esetben, ha az eredeti adathalmazt a tanulók tanulmányi eredményei mentén szétválasztjuk. Mivel az egyes, diszjunkt halmazok mind olyan tanulókat tesznek egy

csoportba, akik hasonló tanulmányi eredményeket értek el, a rájuk vonatkozó determinációs együttható értéke magasabb lesz a kiinduló modell magyarázóerejénél.

Hogy ezt a felvetést ellenőrizni tudjuk, a kiinduló adathalmazt adatbányászati diszjunkt halmazokra, ún. klaszterekre bontjuk.

A klaszterezés egyes típusainak bemutatása

A klaszterezés célja olyan halmazok képzése, melyeknek minden tagja jobban hasonlít a halmaz többi tagjára, mint bármely más halmaz bármely elemére. A klaszterezés típusainak megkülönböztetésére sokféle módszer létezik, a 9. táblázatban Abbas (2008) kutatását alapul véve ismertetem az egyes módszereket és tulajdonságaikat. Négy modellt hasonlított össze, melyeket a legelterjedtebb, legrugalmasabb és legalkalmazhatóbb modelleknek vélt.

Klaszterezési algoritmusok	A megközelítés előnyei
K-mean	Az egyik legelterjedtebb gyorsasága és sorrend-függetlensége miatt. Könnyen megbirkózik a nagyméretű adathalmazokkal is.
Hierarchikus	Könnyen kezeli a hasonlóság és a távolság legtöbb formáját. Minden attribútum-típusra alkalmazható.
Önszerveződő térkép (SOM)	Többféle távolság és illesztési kritérium meghatározható, alkalmazható az attribútumok számának csökkentésére is.
Elvárás-maximalizáló	Jól kezeli a zajos adatokat, erős statisztikai alapokon nyugszik. Jól kezeli a magas dimenziószámot, lineáris a teljesítménye az adathalmaz nagyságával.

10. táblázat: Népszerű klaszterezés algoritmusok és előnyeik (Forrás: saját szerkesztés)

Abbas (2008) kutatása során többek között az alábbi megfigyeléseket tette:

- 1) A K-mean és az Elvárás-maximalizáló algoritmusok teljesítménye nagy elemszámú adathalmaz esetén magasabb, kis elemszámú esetén alacsonyabb a hierarchikus és Önszerveződő térkép algoritmusokhoz képest
- 2) Az Önszerveződő térkép a klaszterek számával arányosan veszít a teljesítményéből
- 3) A K-mean és az Elvárás maximalizáló algoritmusok teljesítménye nagyobb, mint a hierarchikus algoritmusé

- 4) A négy algoritmus közül az Önszerveződő térkép algoritmus teljesítménye átlagosan a legmagasabb

Ezeket a megfigyeléseket figyelembe véve a dolgozat során az Önszerveződő térkép (SOM – Self-Organizing Map) algoritmust alkalmazom, tekintve hogy kis elemszámú adathalmazon kívánjuk alkalmazni a klaszterezést, valamint hogy az erre alkalmas modellek összevetésében ez az algoritmus rendelkezik a legnagyobb átlagos teljesítménnyel.

Klaszterek meghatározása

Jain (2010) kiemeli, hogy a klaszterezés során két fontos dolgot kell meghatároznunk, mielőtt a gyakorlati munkát megkezdénénk. Az egyik legfontosabb eldöntenünk, mi az a szempont, ami alapján a klasztereket elkülönítjük. Erre példaként említi az emlősállatok halmazát. Szét tudjuk választani klaszterekre különféle módon alkothatunk például csoportokat aszerint, hogy az adott faj növényevő, húsevő vagy mindenevő. De besorolhatjuk őket méretük, színük vagy természetes lakhelyük szerint. Ebből jól látszik, hogy csupán a kiindulási szempont mennyire meghatározza a végeredményt.

A másik nagy jelentőséggel bíró tényező, amit Jain (2010) kiemel, hogy mekkorának választjuk a klaszterek számát jelölő k értéket. A legtöbb esetben éppen ezért többféle k értékkel is megvizsgálják az adatokat, és egy előre meghatározott szempont szerint választanak végül ezek közül. Egy ilyen validálási szempontrendszer dolgozott ki Pelleg és Moore (2000), mely többek között a BIC (Bayesian Information Criteria) és a modell komplexitása alapján határozza meg az ideális k számot. Sokszor az adathalmaz eleve rendelkezik egyfajta ideális klaszterszámmal, az emlősöket tekintve például a táplálkozás szerint szétválasztással az ideális szám $k = 3$, hiszen így kapjuk meg a növény-, hús- és mindenevő csoportokat. Ha ugyanezen szempont szerint mondjuk négy csoportra szeretnénk bontani őket, az egyes fajok kevésbé tartoznának össze, ami a validálás során is megmutatkozna.

A modell második szintjén elsőként tehát azt kell eldöntenünk, hogy milyen szempontok szerint érdemes szétválasztani az adatainkat. A válaszhoz újra meg kell vizsgálnunk a dolgozat kérdését.

Fő kérdésünk, hogy milyen összefüggés ismerhető fel a tanulmányi és érettségi eredmények között, és ezt a felismerést hogyan lehet az oktatásban közvetlenül kamatoztatni. A két

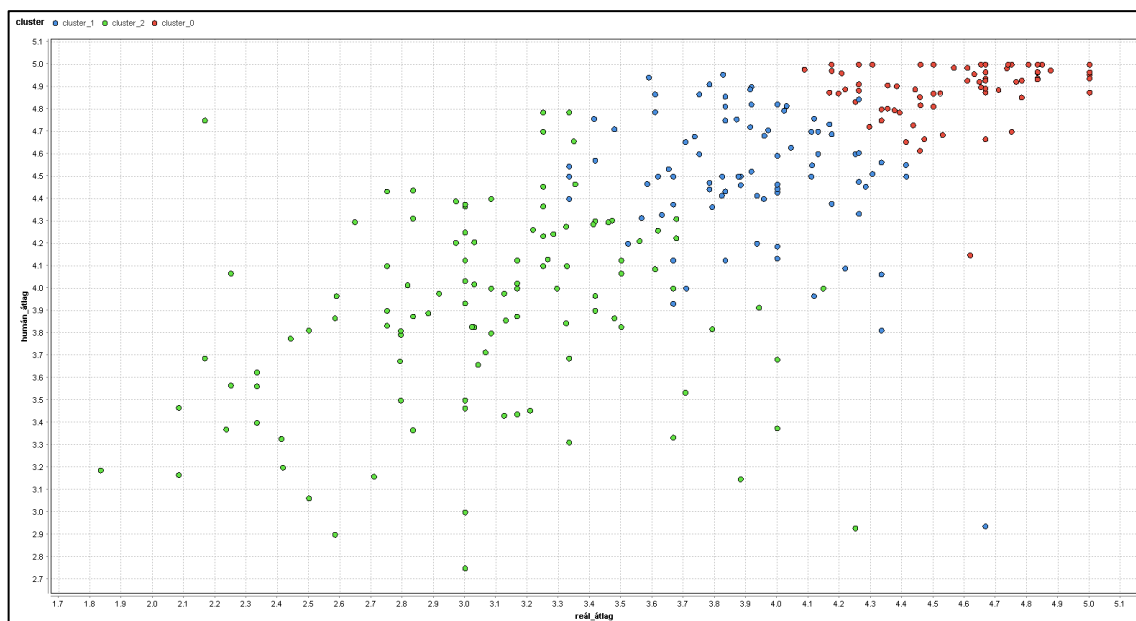
eredménycsoport között a kapcsolat mindig egyirányú, vagyis a tanulmányi eredmények megelőzik az érettségi eredményeket, előbbi értékéből van lehetőségünk következtetni az utóbbiak értékére. Ebből kifolyólag tehát mindenképp a tanulmányi eredmények mentén érdemes szétválasztanunk az egyes klasztereket. A „többletadatok” lehetőséget kínálnak arra, hogy tovább bonthassuk a meglévő adat csoportjainkat.

A szempont kiválasztása után a klaszterek számát kell meghatározni. A szűkített (már csak a választott attribútumokat tartalmazó) adathalmazra a Pelleg–Moore-féle szempontrendszert alkalmazhatjuk, hiszen most már csak az elemzési szempontok mentén választjuk szét az adatokat.

A tanulmányi eredményeket tekintve a folyamat lefutása után a klaszterek ideális számát $k = 3$ -ban határozta meg. Ez alapján a SOM klaszterezési módszert $k = 3$ esetén végrehajtva megkaptuk a modellünk második fázisának alapját képező bontott halmazt (11. ábra).

A 8. ábra alapján a klasztereket az alábbi kategória névvel illelhetjük:

- 1) „Kitűnő tanulók”: Akik mind reál, mind humán szempontból 4,2 feletti átlaggal rendelkeznek (piros klaszter)
- 2) „Jó tanulók”: Humán szempontból átlagos 4,0 felett, reál szempontból 3,5 felett teljesítők (kék klaszter)
- 3) „Nem kiemelkedő tanulók”: Legfeljebb az egyik szempontból érik el a 4,0-s eredményt (zöld klaszter).



11. ábra: Klaszterek a SOM módszer alapján (Forrás: saját szerkesztés)

A modell alkalmazása a klasztereken

Miután a klaszterek mentén szétválasztottuk az eredeti adatokat, 3 új adattáblát nyertünk. Ezekre újra lefuttatjuk az előző lépésben kiválasztott GLM modellt.

A kapott eredményeket a 11. táblázat foglalja össze.

Klaszter	R ² mutató	MSE
„Kitűnő tanulók”	22,8%	2365,27
„Jó tanulók”	40,4%	2122,72
„Nem kiemelkedő tanulók”	35,2%	2403,55
Eredeti GLM modell	48,39%	1349,97

11. táblázat (Forrás: saját szerkesztés)

A klaszterelemzésből levont következtetések

Mindhárom klaszter mutatószámait megfigyelve látható, hogy a klaszterekre vetített modelleknél egyik esetben sem vesznek fel nagyobb értéket, mint a kiinduló modell. Mivel a válogatási szempont nem vezetett a magyarázóerő növeléséhez vagy a hiba csökkenéséhez, kijelenthető, hogy a tanulmányi szempontok szerinti kategorizálása a tanulóknak nem célravezető az érettségi eredmények szempontjából.

A modell eredménye pedig egy nemleges válasz arra a sztereotipizált felvetésre is, miszerint a tanulmányi eredmények nagymértékben előre determinálják az érettségi eredményeket.

8.4 Az osztályzás elemzés alkalmazási lehetőségei az intézményben

Az intézményi alkalmazás infrastrukturális követelményei

Ahogy azt az adattisztítási és adatfeldolgozási folyamatok során is megfogalmaztam, a modellhez szükséges adatforma elérése több nehézségbe is ütközött. Ahhoz, hogy az adott intézményben a jövőben egy hasonló, osztályzat elemzésre alapuló modellt a gyakorlatban alkalmazni lehessen, három kritérium szükséges:

- 1) az e-napló exportálási formátumának sztenderdizálása
- 2) az érettségi adatok digitális rögzítése és megőrzése
- 3) az adatok egy közös adatbázisban való tárolásának lehetősége

A három kritérium teljesülésekor lehetővé válna, hogy az adatok gyors és rendszerezett kinyerésével rövid időn belül felállítható legyen az aktuális adatokra vetített modell.

Becsatornázás az oktatási folyamatba

A modell gyakorlati alkalmazása egyértelműen definiálódik a kiinduló adatok által. Mivel az adott intézményi körülményekre és tanulókra magyarázza a tanulmányi és érettségi eredmények közötti kapcsolatot, alkalmazása is erre a területre vonatkozik.

A modell éves frissítésével és alkalmazásával lehetővé válna egy olyan mutatószám rendszeres kiértékelése, amely közvetlen kapcsolatot mér az intézményi oktatási folyamat és a kimeneti eredmények között. Amennyiben a modell több éves iterációja során beigazolódik annak felhasználhatósága, lehetőség nyílik az egyes tanulók eredményeinek e szempontból történő vizsgálatára is.

9 További osztályzáselemzési lehetőségek

9.1 A jelenlegi modell továbbfejlesztése

A modell elsődleges célja az adott intézményen belül választ találni a tanulmányi és az érettségi eredmények közti összefüggésre, az osztályzat elemzés segítségével. Az adatok időbeli helyzete azonban alapjában korlátozta az elemzés lehetséges mélységét, ahogy azt az adatok aggregálásának szükségességéről szóló részben is kifejtettem.

Amennyiben az elemzés egy olyan környezetben tudna elkészülni, ahol a fent említett korlátozások nem állnak fenn, lehetőség nyílna a modell fejlesztésére.

Faktoranalízis az egyes tantárgyakra

Ha nem szükséges az egyes tantárgyi eredmények összevonása, mélységében megfigyelhetőek lettek volna az egyes tantárgyak érdemjegyeinek hatása a végleges eredményre. Faktoranalízis segítségével kiszűrhetővé válnának a legkevésbé releváns tantárgyi eredmények, így a modell magyarázóerejét is növelni tudnánk.

Idősoros elemzés

Amennyiben több, azonos évekre vonatkozó adattal rendelkező évfolyamot tudunk elemezni, az egyes évfolyamok közti idősoros eltéréseket is be lehet építeni a modellbe, például súlyozással. Az összevetések segítségével az adott évfolyamokra is külön alkalmazható lesz a modell, mely így kevésbé lesz érzékeny az olyan, évfolyam szinten kiugró értékekre, melyek nehezen magyarázhatóak analitika módon (például tanár – vagy osztályfőnök váltás).

Hipotézis-vizsgálatok

A modell továbbfejlesztésének további módja, ha az adatokra vonatkozó, társadalomtudományi szempontból is érdekes kérdéseket megfogalmazó hipotézisekkel bővítjük a modellt. Ilyen például a nemek közötti különbség feltérképezése, az évet ismétlő diákok helyzete, vagy a születési dátum szerinti évfolyamvágás (pl. év vesztesek) potenciális optimalizálása. A modell különböző típusú felállításával vizsgálhatjuk az elmélet gyakorlati oldalát.

9.2 Újabb adatkörök bevonása az osztályzáselemzésbe

Ahogy azt a bevezetőben megfogalmaztam, az általam vizsgált adatkör a jelenleg létező 168 magyar oktatási adatbázisból kettőt kötött össze. A kapcsolatok bővítésére, adatkörök illesztésére rengeteg további lehetőség kínálkozik. Nagy ígéreteket hordoz magában például a kompetencia-felmérések eredményeivel kiegészített tanulmányi adatok elemzése. De nagyon érdekes lenne a sokszor elhanyagolt készség tárgyakkal való összefüggés vizsgálata is, mint például a testnevelés-felmérések, vagy a diákversenyeken elért eredmények.

Ha nagyobb fókuszra választunk, és nem csupán a középiskolát elemezzük, oktatási szinteken átfutó elemzésekre is lehetőség nyílik, így az általános iskolai adatokkal való összekapcsolás például választ adhat arra, hogy a különböző rendszerű középiskolai intézmények (pl. 4-6-8 osztályos gimnáziumok) milyen különböző hatással vannak a tanulókra. Vagy az egyetemi tanulmányi továbbhaladás vizsgálata a középiskolai eredmények tekintetében abban nyújthat segítséget, hogy a középiskolában fontosnak tartott eredmények és mutatók a tanuló későbbi életszakaszában valóban beváltják-e a korábbi reményeket, vagy esetleg más módon lenne érdemes optimalizálni az oktatás menetét és céljait.

Az biztosan állítható, hogy további adatkörök bevonásával egy nagyobb, integrált adatbázis létrehozása is lehetségessé válna. A nagyméretű adathalmazon az analitika pedig felbecsülhetetlen, izgalmas távlatokat nyitó többlet tudás kinyerését tenné lehetővé.

10 Összefoglalás

Ahogy azt a bevezetőben is megfogalmaztam, az oktatás fejlődése már évtizedek óta fontos téma, reformja pedig a küszöbön áll. Ennek egy megoldását sokan, köztük Baker és Siemens (2014) is, az oktatás-analitika fejlődésében és átfogó alkalmazásában látják.

Bemutattuk, hogy ennek egy alulról szerveződő folyamatnak kell lennie, és mint ilyen, az egyes intézményekben történő adatelemzéseknek köszönhetően fog elterjedni, majd a gyakorlatban is hasznossá válni. Dolgozatom célja pontosan ez volt: az választott intézmény számára az oktatási elemzések segítségével többlet tudást nyújtani az ott végbemenő tanulmányi folyamatokról.

Az elemzés szintjének az oktatási adatbányászatot, azon belül is az osztályzat elemzést választottam, ennek keretein belül kerestem arra a választ, milyen kapcsolatot találunk az intézményben a kinyerhető tanulmányi adatok és az érettségi eredmények között.

A kérdést kétféleképpen válaszoltam meg: egy felállított prediktív modell segítségével, valamint a modell egy gyakorlati alkalmazásának bemutatásával.

Dolgozatom során az adatbányászatban elterjedt CRISP-DM módszertan alapján állítottam fel a prediktív modellt. A módszertan lépésein végighaladva előbb a környezet megértését, majd az adatok bemutatását, tisztítását és elemzési formába öntését mutattam be. A következő fázisban részleteztem a modell felállítását, majd értékeltem a kapott eredményeket. A folyamat kimeneteként egy 48,39%-os magyarázó erejű prediktív modell állítottam fel.

Az értékelésekor figyelembe vettem az adatok nem reprezentatív voltát azok speciális származási körülményei miatt, valamint a minta kis elemszámából, és az adatkinyerés helyzeti nehézségeiből fakadó alapvető mérési és feldolgozási pontatlanságokat és korlátokat.

Emellett a dolgozat részeként bemutattam a modell gyakorlati alkalmazási lehetőségeinek egy példáját is. Azt a sokfelé elterjedt, intuitív elképzelést vizsgáltam, miszerint a jobb tanulmányi eredménnyel rendelkező tanulók jobb érettségi eredményt, míg a gyengébb eredményeket felmutató tanulók rosszabb érettségi eredményt érnek el.

Klaszterező módszerrel kisebb egységekre bontva a kiinduló adathalmazt bemutattam, hogy az egyes klaszterekre képzett ugyanolyan modell magyarázó ereje minden esetben kisebb lett, mint a teljes egészként kezelt adathalmazra vonatkoztatott magyarázó erő. Ebből egyértelműen levonható a következtetés, hogy ez a fajta klaszterezés nem finomítja a modellt, vagyis az eredeti adathalmaznak sem volt természetes sűrűsödése a választott szempont szerint. Azaz az eredeti, intuitív elképzelés a modell által elutasításra került.

Összességében kijelenthető, hogy bár a modell gyakorlati előrejelzésre statisztikai szempontból nem alkalmas, ugyanakkor a dolgozat vizsgált kérdéseire értékelhető és értelmezhető választ adott, továbbá alátámasztja azt is, hogy az osztályzáselemzés koncepciójának gyakorlatba ültetésével érdemes foglalkozni.

Irodalomjegyzék

- Abbas, A. O. (2008). *Compararison Between Data Clustering Algorithms*. The International Arab Journal of Information Technology, 2008. Vol 5, No 3. pp. 320-325.
- Baker, R. S. J. D. (2010). *Data Mining For Education*. In International Encyclopedia of Education (3rd edition), Elsevier, Oxford, UK.
- Baker, R. S. J. D. – Yacef, K. (2009). *The State of Educational Data Mining in 2009: A Review and Future Visions*. Journal of Educational Data Mining, Article 1, Vol 1, No 1, Fall 2009.
- Baker, R – Siemens, G. (2014). *Educational Data Mining and Learning Analytics*. pp. 253-272.
- Felvi.hu (2018) *Felsőiskolás felvételi tájékoztató, 2018-ban induló képzések*. https://www.felvi.hu/felveteli/jelentkezes/felveteli_tajekoztato/FFT_2018A/2_pontszamitas (Letöltve: 2018. április 3.)
- Fodor Szabolcs – Horváth, Tamás – Köblös István, Szemenyei Mariann. (2016). *Tanulói életpálya nyomkövetés*. Digitális jólét program. 2016 június.
- Frochte, J. – Bernst, I. (2016). *Success Prediction Systems For Student Counseling Using Data Mining*. 8th Conference on Knowledge Discovery and Information Retrieval, Porto, Portugal. November, 2016.
- Hershkovitz, A., – Nachmias, R. (2008) *Developing a Log-Based Motivation Measuring Tool*. Proceedings of the 1st International Conference on Educational Data Mining, pp. 226-233.
- Hunyadi László – Vita László (1991). *Statisztika I*. Aula Kiadó, Budapest, 1991.
- Hunyadi László – Vita László (1992). *Statisztika II*. Aula Kiadó, Budapest, 1992.
- Jain, A. K. (2010). *Data Clustering: 50 years Beyond K-means*. Pattern Recognition Letters, 2010., Vol 31, Issue 8, pp. 651-666.

- Kohavi, R. (1995). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. In: International Joint Conference on Artificial Intelligence, 1995, August. Vol. 14. No.2, pp.1137-1145.
- Liu, Y. – Li, Z. – Xiong, H. – Gao, X., Wu, J. (2010). *Understanding of Internal Clustering Validation Measures*. IEEE Conference on Data Mining 2010., pp. 911-916.
- Long, P. – Siemens, G. (2011). *Penetrating the Fog: Analytics in Learning and Education*. EDUCAUSE Review, September/October 2011., pp. 31-40.
- Madhulatha, T. S. (2012). *An Overview on Clustering Methods*. IOSR Journal of Engineering. April 2012, Vol. 2(4), pp. 719-725.
- Milligan, W. G. – Cooper, M. C. (1987). *Methodology Review: Clustering Methods*. Applied Psychological Measurement 1987., pp. 329-354.
- Molnár Gyöngyvér (2011). *Az információs-kommunikációs technológiák hatása a tanulásra és oktatásra*. Magyar Tudomány, 2011/9. szám, 1038-1046. oldal.
- Oster, D. W. (1988). *Selection of Optimal Regression Models via Cross-Validation*. Journal of Chemometrics, 1988., Vol. 2., pp39-48.
- Pelleg, D., – Moore, A. W. (2000). *X-means: Extending K-means with Efficient Estimation of the Number of Clusters*. Proceedings of the 17th International Conference on Machine Learning, 2000., pp.727-734.
- RapidMiner (2018). *RapidMiner Documentation*. RapidMiner GmbH 2018. <https://docs.rapidminer.com/latest/studio/operators/> (Letöltve: 2018.04.04.)
- Romero, C. – Ventura, S. (2007). *Educational Data Mining: A survey from 1995 to 2005*. In: Expert Systems with Applications 33, 2007., pp135-146.
- Sclater, N. – Mullan, J. (2017). *Learning Analytics and Student Success – Accessing the Evidence*. JISC Briefing, January 2017.

Siemens, G. – Gasevic, D. – Haythornwaite, C. – Dawson, S. – Shum, S. B. – Ferguson, R. – Duval, E. – Verbert, K. – Baker, R. S. J. D. (2011). *Open Learning Analytics: an Integrated & Modelarized Platform*. SoLAR, July 28, 2011.

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. April, 2000., pp29-39.

Witten, I. H. – Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kauffman, San Fransisco, CA.

Zielke, T. – Bernst, I. – Crombach, T. – Becker, M. – Lippel, J. (2018). *What Can Be Learned by Grading Analytics?* Proceedings of INTED2018 Conference, Valencia, Spain. 5th-7th March, 2018., pp3097-3106.

1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, February 27–March 1, 2011, <https://tekri.athabascau.ca/analytics/> (Letöltve: 2018.03.12.)

Ábrajegyzék

1. ábra: A CRISP-DM életciklus	10
2. ábra: Oktatással kapcsolatos adatbázisok Magyarországon	15
3. ábra: Tanulmányi eredmények egy nyers adattáblája.....	24
4. ábra: Sztenderdizált adattábla	25
5. ábra: Nyers érettségi adattábla	26
6. ábra: Érettségi eredmények egységesítése.....	26
7. ábra: Véglegesített, illesztett adattábla	29
8. ábra: RapidMiner modell struktúra létrehozása.....	32
9. ábra: RapidMiner modell adatok felvétele	33
10. ábra: RapidMiner modell eredmények	33
11. ábra: Klaszterek a SOM módszer alapján.....	45