Adatbányászati módszerek alkalmazása oktatási környezetben keletkezett adatokon

Szücs Katalin[1], Kiss Attila[2]

[1]kszucs@caesar.elte.hu
ELTE IK
[2]kiss@inf.elte.hu
ELTE IK

Absztrakt. Az adatbányászat célja nagy adattömegek rejtett információjának feltárása, a triviálisan nem észlelhető mintázatok és összefüggések kinyerése. Eszközei a legkülönbözőbb területeken kerülnek felhasználásra. A tanítási környezetben elért sikeres alkalmazások egy új kutatási irány, az oktatási adatbányászat kialakulásához vezettek. A megközelítés célja középiskolák, egyetemek, intelligens tanító rendszerek által előállított adatokból hasznos információ előállítása adatbányászati módszerek alkalmazásával, annak érdekében, hogy választ kapjunk a diákok tanulási folyamatával kapcsolatos kutatási kérdésekre, valamint támogatást nyújtsunk a különböző tanítási módszertanok fejlesztéséhez.
Kutatásunk célja az adatbányászat ilyen irányú alkalmazásainak összefoglalása a szakirodalom alapján.

Kulcsszavak: adatbányászat, oktatási adatbányászat, digitális oktatási rendszerek.

1. Bevezetés

Az oktatási adatbányászat fiatal kutatási terület, 2006-ban jelentek meg az első publikációk a témában. [7] A különböző képzések során begyűjtött statisztikák, a rögzített felmérési eredmények és az egyre nagyobb teret hódító interaktív tanulási környezetek elegendő adatot szolgáltatnak ahhoz, hogy az adatbányászat módszereit oktatási területen is hasznosítani tudjuk. Az online tananyagok interaktív tanulása, tesztek kitöltése során olyan adatok keletkeznek nagy mennyiségben, amelyek eddig nem álltak rendelkezésre. Mint például, hogy mennyi időt tölt egy diák az egyes kérdések megválaszolásával, vagy milyen segítséget vesz igénybe. Ezekből az adatokból kinyerhető, hogy hogyan dolgozza fel az anyagot, hogyan tanul a tanuló. Mivel ezen ismeretek segítségével számos eddig kevéssé körvonalazott részletbe nyerhetünk betekintést a diákok tanulási folyamatát illetően, az új megközelítés nagy mértékben hozzájárulhat a különböző képzések színvonalának növeléséhez, hatékonyabb módszertanok kialakításához. Az újonnan elérhető adatok adatbányászati módszerekkel való feldolgozása után többek között olyan kérdésekre kaphatunk választ, mint, hogy hogyan készítsünk hatékonyabb oktatási segédanyagot, hogyan jósolhatjuk meg a tanulók jövőbeli teljesítményét, vagy miként adhatunk számukra megbízható tanácsot szakirányválasztáskor.

A digitális tanulási környezet megadja a lehetőséget, hogy az oktatók folyamatosan és részletesen megfigyelhessék a diákok viselkedését a tanulás teljes folyamata során. Az oktatási rendszerek alapvető statisztikákat szolgáltatnak a tanulási szokásokról, ilyen információ lehet például a rendszerbe való belépések gyakorisága, a különböző látogatások hossza, a sikeresen megoldott feladatok mennyisége, vagy a különböző feladatok megoldása során igénybe vett segítségek száma. A rendszer az összes elérhető adatot log fájlokban rögzíti. Ezek a statisztikák viszont önmagukban csupán távoli következtetéseket engednek levonni a tanulási folyamat pontos menetéről. A log fájlokban tárolt nyers adatok adatbányászati módszerekkel történő elemzése után

lehetővé válik olyan implicit információ felfedezése a diákok aktivitásáról, ami megmutatja, mik a tipikus viselkedési minták és ezek közül melyek azok, amelyek várhatóan nagyobb sikert eredményeznek a tanulási folyamat végén. Az ilyen típusú vizsgálatok elsődleges célja a tanulási folyamat testreszabása a tanuló számára, hogy adaptív visszajelzéseken és személyre szabott értékeléseken keresztül megállapíthassák, hogy mi az a tanulási út, amelyet végigjárva az adott diák teljesítménye maximalizálható, valamint melyek azok az indikátorok, amelyek a tanulás minőségét leginkább meghatározzák.

Az oktatási rendszerek működése több szempontból is vizsgálható az adatbányászat eszközeivel. Például különbséget kell tennünk a tananyag tartalmi és szervezési oldalról történő vizsgálata között. [13] Az információ kinyerése az első esetben általában szöveganalitikai eszközökkel történik. A szervezési oldalról történő vizsgálat célja a tananyag hatékonyságának elemzése, itt a webes alapú digitális oktatási rendszerek által rögzített adatok elemzése kerül a fókuszba. Annak ellenére, hogy munkánk további részében a diákok tanulási folyamatának vizsgálatára kerül a hangsúly, számos más céllal is érdemes lehet oktatási adatbányászatot végezni. Tananyag fejlesztési szempontból például segítséget jelent, ha tudjuk, milyen sorrendben építsük egymásra a tananyaghoz tartozó különböző elemeket, vagy választ kapunk olyan kérdésekre, mint, hogy mitől számít nehéznek egy témakör, vagy melyek a hasonló nehézségű feladatok.

2. A tanulási folyamat megértése

A szakirodalomban számos kutatással találkozhatunk, amelyek adatbányászati eszközöket alkalmaznak oktatási környezetben keletkezett adatokon. A vizsgálatok célja olyan hasznos információ kinyerése az adatokból, amelyek mélyebb betekintést engednek a tanulási folyamatba, így előnyt jelentenek a különböző didaktikai módszerek tervezése során. Most ezekből az eredményekből mutatunk be néhányat.

2.1. Tanulási folyamat vizsgálata digitális oktatási rendszerekben

Ebben az alfejezetben néhány példát mutatunk arról, hogy milyen új ismereteket szerezhetünk a diákok tanulási folyamatáról, ha a digitális oktatási rendszer által gyűjtött adatokat adatbányászati módszerekkel elemzünk. Részletesen ismertetünk néhány fontos adatbányászati módszert, amelyek alkalmazása egy tajvani kutatás eredményein kerül bemutatásra. [4]

A kutatás 98 tajvani diák viselkedését elemezte egy online kurzus elvégzése során. Az adatok begyűjtése a Wisdom Mater v.2.4 rendszeren keresztül történt, ami az egyik legnépszerűbb digitális oktatási keretrendszer Tajvanban. A vizsgálat célja a tanulók tipikus online tanulási viselkedésének megismerése, a leggyakoribb viselkedési minták és az ezeket befolyásoló legfontosabb tényezők feltárása volt. A kutatáshoz felhasznált adatok egy hathetes intervallumot ölelnek fel, aminek során összesen 17934 szerver log került rögzítésre.

Változónév	Leírás
ID	Felhasználói azonosító
LoginFre	Bejelentkezések száma
LastLog	Utolsó bejelentkezés időpontja
ClassFre	A tananyaghoz való hozzáférések gyakorisága
LastClass	Az utolsó hozzáférés időpontja
NoPosting	A faliújságra küldött üzenetek száma
DisFre	Párbeszédekben való részvételek száma
ReadHr	A faliújság olvasásával eltöltött idő

ReadMegs	A faliújságon elolvasott üzenetek száma

1. táblázat: Az elemzésben felhasznált változók listája

Az 1. táblázat tartalmazza a logokból kinyert változók nevét és leírását. Az így előállított adatokat leíró, valamint prediktív vizsgálatnak is alávetették.

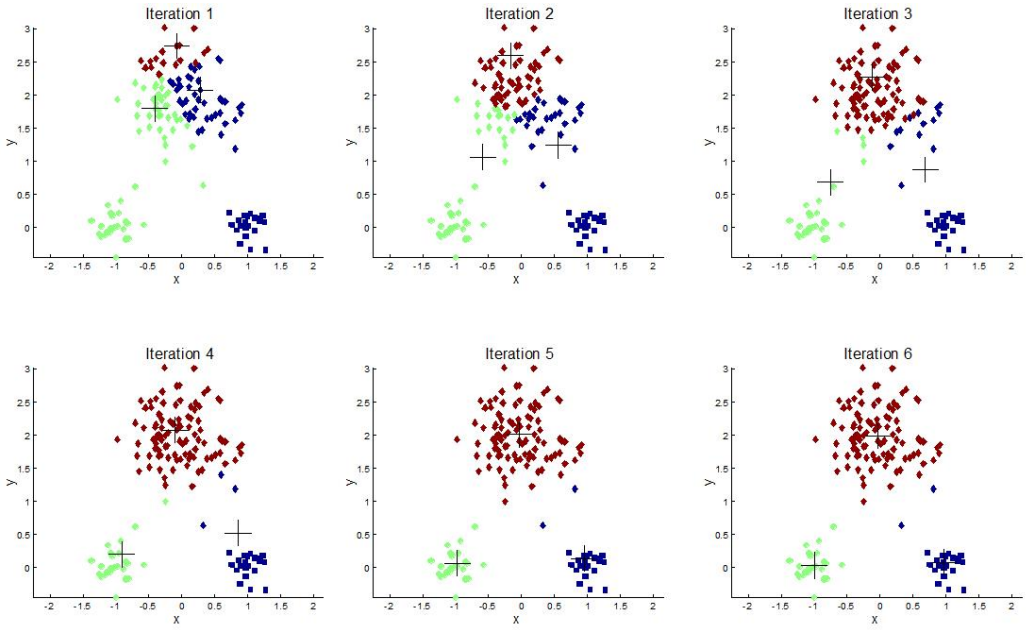
A leíró vizsgálat első lépésében a k-közép algoritmus segítségével csoportosításra kerültek a hasonló tulajdonságokkal rendelkező diákok. Az algoritmus általános célja n darab megfigyelés k klaszterbe való sorolása olyan módon, hogy minden megfigyelés a hozzá legközelebb eső klaszterközéppontú klaszterben kapjon helyet. Lépései, melyeket az 1. ábra[1] példaadatsoron mutat be, a következők:

lépés: kiválasztunk k darab megfigyelést véletlenszerűen. Ezek alkotják kezdetben a klaszterközéppontokat.

lépés: minden megfigyelést a hozzá legközelebb eső klaszterközépponthoz rendelünk.

lépés: helyettesítsünk minden klaszterközéppontot a klaszterhez tartozó elemek átlagával.

lépés: ismételjük a 2-es és a 3-as lépést. Ha a lépések elvégzése után nincs több változás, az algoritmus megáll.



1. ábra: A k-közép algoritmus működése példaadatsoron.

A klaszterezés eredményét a 2. táblázat foglalja össze. A vizsgálat során a táblázatban feltüntetett hat változó került felhasználásra. Az eljárás a hasonló tulajdonságokkal rendelkező diákok három diszjunkt halmazát határozta meg, az 1-3. oszlopok ezen halmazok klaszterközéppontjait tartalmazzák. A 4. oszlop pedig a teljes adathalmazra vonatkozó átlagokat mutatja. Az 1. és a 2. klaszterbe sorolt diákok teljesítménye meghaladja az átlagot. Jól látható, hogy az ebbe a két cso-

1 Kép forrása: https://apandre.wordpress.com/visible-data/cluster-analysis/

portba tartozó 36 diák minden viselkedési változón is magasabb értékeket értek el, mint az átlag. Érdekes megfigyelni, hogy annak ellenére, hogy a 2. klaszterbe tartozó tanulók érték el a legmagasabb eredményeket, a viselkedési változók átlagos értékei itt alacsonyabbak, mint az első klaszter esetében. Ez arra enged következtetni, hogy a 2. klaszterbe tartozó tanulók felkészülése hatékonyabb volt. A 3. klaszterben található 62 diák teljesítménye az átlagnál gyengébb és az is megfigyelhető, hogy a különböző változók értéke is elmarad az átlagostól, ami azt jelenti, hogy ők kevesebb aktivitást mutattak a kurzus elvégzése során.

Változónév	1. Klaszter N=29	2. Klaszter N=7	3. Klaszter N=62	4. Klaszter N=98
Bejelentkezések száma	50.83	49.86	23.65	33.56
A tananyaghoz való hozzáférések gyakorisága	53.48	51.86	24.06	34.76
A faliújságra küldött üzenetek száma	57.04	40.71	16.35	30.13
A faliújságon elolvasott üzenetek száma	91.86	24.86	24.11	44.21
Párbeszédekben való részvételek száma	3.97	2.00	1.63	2.35
Eredmény	83.13	87.13	63.32	70.60

2. táblázat: A teljes adathalmaz klaszterezésének eredménye.

A k-közép algoritmus alkalmazásával ezt a klasztert tovább bontották és két újabb al-klaszter kerültek meghatározásra, ennek eredményét a 3. táblázat foglalja össze. Jól látható, hogy így a 3-1. klaszterbe csoportosított 44 diák a csoport átlagával arányos teljesítményt ért el, ugyanakkor a 3-2. klaszterhez tartozó 18 diák ritkán látogatta az online felületet és az eredményük végül nagyban elmaradt a csoport átlagától.

Változónév	3-1. Klaszter N=44	3-2. Klaszter N=18	3. Klaszter N=62
Bejelentkezések száma	31.60	4.23	23.65
A tananyaghoz való hozzáférések gyakorisága	32.45	3.56	24.06
A faliújságra küldött üzenetek száma	20.42	6.40	16.35
A faliújságon elolvasott üzenetek száma	31.05	7.17	24.11
Párbeszédekben való részvételek száma	2.00	0.72	1.63
Eredmény	71.18	44.11	63.32

3. táblázat: A gyengébben teljesítő tanulók klaszterezésének eredménye.

A leíró elemzés második lépésében a változók halmazára vonatkozóan asszociációs szabályok kerültek meghatározásra, hogy feltárják a közöttük megfigyelhető kapcsolatokat és így következtetéseket vonhassanak le a tanulók viselkedéséről. Az asszociációs szabályok a változók különböző részhalmazainak együtt előfordulásáról adnak információt és a következő alakban kerülnek

meghatározásra: „viselkedés A → viselkedés B, támogatottság = 32%, megbízhatóság = 80%”. A szabály támogatottsága azt jelzi, hogy az összes megfigyelt esemény hányad részében figyelhető meg együtt mind az A, mind a B viselkedés. A megbízhatóság pedig a B viselkedés előfordulásának valószínűsége egy olyan megfigyelés során, amelyben szerepel A.

A 4. táblázat a leggyakoribb napi aktivitásokhoz tartozó asszociációs szabályokat tartalmazza. Az első szabály alapján megállapítható, hogy a napi tevékenységek több, mint 50%-át a tananyag olvasása jelenti. A 3. szabály alapján az is megállapítható, hogy ha a diákok nekiláttak a tananyag olvasásának, akkor több, mint 57% eséllyel még egyszer megnyitották az anyagot a nap folyamán. Hasonlóképpen a 4. szabályból látható, hogy ha a nap során megtörtént egy bejelentkezés, 46.67% valószínűséggel történt még bejelentkezés az adott diáktól aznap.

Szabály	Támogatottság	Megbízhatóság
1. Belépés a tanulási környezetbe → tananyag olvasása	57.06	90.14
2. Sikeres bejelentkezés → tananyag olvasás	40.15	70.65
3. Tananyag olvasás → tananyag olvasása	34.05	57.59
4. Sikeres bejelentkezés → sikeres bejelentkezés	26.53	46.67
5. Belépés a tanulási környezetbe → üzenet küldése a faliújságra	30.24	40.28
6. Üzenet küldése a faliújságra → üzenet küldése a faliújságra	20.69	73.53

4. táblázat: A legfontosabb asszociációs szabályok.

Ezt követően egy döntési fa került megépítésre az adathalmaz prediktív vizsgálatának érdekében. A vizsgálat célja a diákok eredményét befolyásoló legfontosabb indikátorok meghatározása annak érdekében, hogy egy esetleges új diák végeredménye kellő pontossággal megjósolható legyen csupán a többi változó ismeretében.

A döntési fák elmélete az adatbányászat témakörébe tartozik. Széles körben alkalmazott módszer, számos döntéstámogató rendszer alapeleme. Lényege, hogy az attribútumok egy olyan fa struktúrába szervezett rendszerét hozzuk létre, amelynek segítségével egy újonnan érkező elemről egyszerűen eldönthető a célváltozó legvalószínűbb értéke. A fa gyökere és a köztes csúcsok az attribútumokra vonatkozó feltételeket tartalmaznak, a fa levelei pedig a célváltozóra vonatkozó döntéseket. Egy új elemet a fa gyökerétől indítva az attribútumokra vonatkozó döntések egyértelműen vezetik végig a köztes csúcsokon, amíg az el nem jut egy levél csúcsba mely alapján meghatározásra kerül a célváltozó értéke. Ha a megfelelő módon konstruált döntési fa már rendelkezésre áll, egy új elem kiértékelése egyszerű. Az optimális fa meghatározása azonban a legtöbb esetben számításilag kivitelezhetetlen. A Hunt algoritmus [15] viszont lehetővé teszi olyan döntési fák konstruálását elfogadható időn belül, amelyek közel optimális eredménnyel teljesítenek. Ez a rekurzív és mohó algoritmus a következő módon működik. Tegyük fel, hogy D_t jelöli a fa építése során a tanuló adathalmaz t csúcsba érkező részhalmazát, $y = \{y_1, \ldots, y_c\}$ vektor pedig a célértékeket tartalmazza.

1. lépés: ha D_t minden eleme ugyanahhoz a célértékhez, y_t-hez tartozik, akkor legyen t a döntési fa egy levele y_t címkével.

2. lépés: ha D_t különböző célértékekhez tartozó elemeket tartalmaz, akkor határozzunk meg egy attribútum tesztfeltételt, amellyel az adatokat kisebb részhalmazokra bontjuk, oly módon, hogy a keletkező halmazokban a címkék diverzitása minél alacsonyabb legyen. Az így létrejött halmazok mindegyikéhez egy új csúcsot hozunk létre a t csúcs gyerekeként. Az így létrejött csúcsokban az eljárást rekurzív módon folytatjuk.

A diákok adatainak döntési fával történő elemzése során kiderült, hogy a végeredményt befolyásoló legfontosabb attribútum a tananyaghoz való hozzáférés gyakorisága. Azok a diákok, akik a hat hét alatt több, mint 18,5-szer nyitották meg a tananyagot átlagosan 77.92%-os eredménnyel végeztek, azok pedig, akik több, mint 44.5-szer, 89.62%-os átlagot teljesítettek. A második legfontosabb változóként a faliújságon elolvasott üzenetek száma került ki. Azok a diákok, akik 66.5 üzenetnél többet olvastak, 15.43%-al az átlag fölött teljesítettek.

2.2. Teljesítmény előrejelzés a Random Forest algoritmussal

A KDD Cup egy évente megrendezésre kerülő nemzetközi adatbányász verseny. 2010-ben a verseny témája oktatási adatbányászat volt. [10] A szervezők több digitális oktatási rendszer által rögzített adatot tettek elérhetővé a résztvevők számára. Ebben az alfejezetben azt mutatjuk be, hogy ezek az oktatási rendszerek pontosan milyen típusú adatokat képesek szolgáltatni, és hogy ezek az adatok hogyan használhatóak fel a diákok teljesítményének előrejelzésére.

Az adatok két forrása egyfelől a Carnegie Learning Algebra 2005-2006 és 2006-2007 között rögzített adatai, másrészt pedig a Bridge to Algebra nevű rendszer 2006-2007 közötti adatai. Az oktatási rendszerek a diákok munkáját részfeladat megoldásonként rögzítették, így a keletkezett adatbázis minden sora egy adott diák egy bizonyos részfeladat megoldásához tartozik. Az adatbázis azt az információt is tartalmazza, hogy mely feladatok teljesítéséhez milyen tudásanyag előzetes ismeretére van szükség. Néhány fontos attribútum, ami a versenyzők rendelkezésére állt a diákok tanulási rendszerben végzett aktivitásáról:

1. A diák azonosítószáma
2. A feladat címe
3. A feladat elhelyezése a tantervi hierarchiában
4. A feladat azonosítószáma
5. A diák összesen hányszor találkozott az adott feladattal
6. A részfeladat azonosítója
7. A részfeladat elkezdésének ideje
8. A részfeladatra történő első válaszadás ideje
9. A részfeladatra történő első helyes válaszadás ideje
10. A részfeladatra történő utolsó válaszadás ideje
11. Az első és az utolsó válaszküldés idejének különbsége
12. Hibás válaszadási kísérletek száma
13. Helyes válaszadási kísérletek száma
14. A részfeladat megoldása során felhasznált segítségek száma
15. A feladat megoldásához szükséges előismeretek listája
16. Azon részfeladatok száma, amelyekkel a diák már foglalkozott és ugyanarra a tudás anyagra épül, mint az aktuális feladat.
17. Sikeres volt-e a részfeladat megoldása az első próbálkozás során: 1, ha igen, 0, ha nem (CFA).

A résztvevők számára az adatokat két diszjunkt részhalmazra osztották, ezek alkották a tanuló és a teszt adathalmazokat. Minden diákhoz véletlenszerűen kiválasztották az egyik feladatmegoldás adatait és a teszt halmazhoz adták. Azokhoz a feladatokhoz tartozó sorokat, amiket a diák a kiválasztott feladat előtt oldott meg, a tanuló halmazhoz adták, a többi sort, vagyis a kiválasztott feladat után megoldott feladatok sorait pedig elvetették. A tanuló halmaz a fent említett minden

attribútumot tartalmazta, a teszt halmazból azonban hiányoztak a diákok teljesítményére vonatkozó változók, vagyis a fenti lista 8-11. és 17. elemei.

A feladat egy prediktív modell építése volt a tanuló adathalmaz alapján, aminek alkalmazásával megjósolható a teszt halmazban található részfeladatok teljesítésének eredménye. Pontosabban a CFA változó értékének minél pontosabb meghatározása volt a cél.

A legjobb eredményt a versenyzők a Random Forest nevű eljárás alkalmazásával érték el. [16,17] Ez a megközelítés az úgy nevezett együttes (ensemble) vagy kombinált osztályozók családjába tartozik. Egy együttes módszer alaposztályozók egy halmazát hozza létre a tanulóadatokból, hogy ezek egyszerre történő alkalmazásával jobb előrejelzést érjen el, mint bármelyik alaposztályozó önálló használatával. A módszer fő gondolata, hogy „gyengébben tanuló" elemek csoportjából egy sikeres osztályozót állít össze. A Random Forest osztályozó alaposztályozói döntési fák, így ebben a kontextusban ezek alkotják a „gyengébben tanuló" elemek csoportját. A tanítási fázisban minden döntési fa külön épül a tanuló adathalmaz egy-egy véletlenszerűen (visszatevéssel) választott részhalmazán. A fa minden csúcsában az attribútumokból m darab kerül kiválasztásra véletlenszerűen, ezek közül kerül ki az, amelyik felhasználásával a legjobb tesztfeltételt határozhatjuk meg a csúcsba beérkezett adatok kettéválasztására. A vágást meghatározó attribútum jelölteknek ezt a véletlen módon történő kiválasztását nevezik baggingnek. Egy új elem kiértékelésének során először a célváltozó értéke egyesével meghatározásra kerül minden döntési fában, majd a végkimenetelek értékéből a végső eredmény átlagolással, vagy kategorikus attribútumok esetén a többségi döntés elve alapján számítódik ki.

A versenyzők által küldött eredmények összehasonlítása a modellekhez tartozó négyzetes középhiba (RMSE) alapján történt, ami nevéhez híven a megfigyelt és a predikált értékek közti különbségek egy mérőszáma .A legjobb megoldáshoz tartozó RMSE érték 0.2815 volt.

2.3. Teljesítmény előrejelzés ajánlórendszerrel

A diákok teljesítményének előrejelzése számos előnnyel járhat egy digitális oktatási rendszerben. Lehetőséget ad például a feladatok megfelelő sorrendjének kiválasztására, a segítségnyújtás vagy figyelmeztetés megfelelő pillanatának meghatározására, vagy a megfelelő nehézségi szint beállítására. Azt, hogy egy diák várhatóan hogyan teljesít egy bizonyos feladatot, meghatározható ajánlórendszerek segítségével is. [18]

Az ajánlórendszereket hagyományosan arra használják, hogy nagy mennyiségű adatból kiszűrjék a leghasznosabb elemeket. Az egyik leghíresebb ilyen rendszer az Amazon ajánlórendszere, ami egy adott termékhez megjósolja, hogy egy vevő számára az mennyire lehet értékes, a vevő által más termékekre adott értékelések alapján. Ha ugyanezt az eljárást a KDD Cup 2010 verseny megoldására szeretnénk használni, megfeleltethetjük a vásárlókat a diákokkal, a feladatokat a termékekkel, a termék értékelést pedig a CFA értékével. Így az ajánlórendszer alkalmazásával megjósolhatjuk, hogy sikeresen oldja-e meg az adott részfeladatot a tanuló, az alapján, hogy más, hasonló tanulási utat bejárt diákok hogyan teljesítették a részfeladatot.

4. Összefoglalás

Munkánk során az adatbányászati módszerek oktatási környezetben való integrálhatóságát vizsgáltuk, nagy hangsúlyt fektetve a diákok tanulási folyamatának megismerésére. Példákon keresztül tekintettük át az adatbányászat ilyen irányú alkalmazásait. Megmutattuk, hogy a digitális oktatási környezetből kinyert adatokon hogyan lehet például klaszterezés segítségével megtalálni a hasonló módon tanuló diákok halmazát. Arra is példát mutattunk, miként határozhatóak meg a diákok teljesítményét legnagyobb mértékben befolyásoló faktorok. Végül mutattuk, hogyan jó-

solható meg a diákok jövőbeli teljesítménye múltbeli adatok alapján a döntési fák, a Random Forest és ajánló rendszerek alkalmazásával.

A hasonló sikeres alkalmazások száma a jövőben várhatóan tovább bővül a digitális oktatási rendszerek egyre nagyobb népszerűségének köszönhetően. Ezért az adatbányászat eszközeinek alkalmazása az oktatásban továbbra is aktív kutatási terület.

Irodalom

Educational data mining: http://www.educationaldatamining.org/

Heiner, Cecily, Neil Heffernan, and Tiffany Barnes. "Educational data mining." *Supplementary Proceedings of the 12th International Conference of Artificial Intelligence in Education*. 2007.

Erika, Jókai, Horváth Cz János, and Horváth Ádám. "MOODLE alapú tantárgyat végző hallgatók tanulási szokásainak elemzése adatbányászati eszközökkel."

Hung, Jui-Long, and Ke Zhang. "Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching." *MERLOT Journal of Online Learning and Teaching* (2008).

KDD Cup 2010: https://pslcdatashop.web.cmu.edu/KDDCup/rules_data_format.jsp

P. Cortez and A. Silva. "Using Data Mining to Predict Secondary School Student Performance." *In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.*

deLaFayette Winters, Titus. „Educational data mining: collection and analysis of score matrices for outcomes-based assessment." Diss. University of California Riverside, 2006.

László, Bóta. „Oktatási adatbányászat" Agria Media 2011

Stamper, J., Niculescu-Mizil, A., Ritter, S., Gordon, G.J., & Koedinger, K.R. (2010). Algebra I 2008-2009. Challenge data set from KDD Cup 2010 Educational Data Mining Challenge. Find it at http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp

Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1.1 (1986): 81-106.

Feature engineering and classifier ensemble for KDD cup 2010. *KDD Cup*, 2010.

Svetnik, Vladimir, et al. "Random forest: a classification and regression tool for compound classification and QSAR modeling." *Journal of chemical information and computer sciences* 43.6 (2003): 1947-1958.

Thai-Nghe, Nguyen, et al. "Recommender system for predicting student performance." *Procedia Computer Science* 1.2 (2010): 2811-2819.