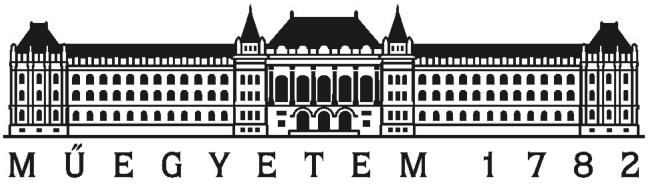


TDK DOLGOZAT

Elsőéves hallgatók pandémia előtti és alatti bemeneti adatainak elemzése modern adattudományi eszközökkel

Köller Donát Ákos & Vlaszov Artúr
BME Matematikus MSc

Témavezető: Szilágyi Brigitta
Geometria Tanszék

BME Matematika Intézet
Budapest
2022

Tartalomjegyzék

1. Bevezetés

Az elsőévesek képességének és készségeinek mérése nemcsak a felsőoktatásban bevett gyakorlat, hasonlóval találkozhatunk akkor is, amikor például a diákok a gimnáziumi tanulmányaikat kezdik meg, de bevett gyakorlat a szintfelmérő a nyelviskolákban is a nyelvi szintek megállapításánál. A felsőoktatásban leginkább arra használják a bemeneti méréseket, hogy feltérképezzék, a hallgató birtokában van-e az adott tárgy teljesítéséhez szükséges előismereteknek. Amennyiben nincs, felzárkóztató kurzusokat biztosítanak. Előfordulhat az is, hogy valamely egyetemi tárgy teljesítése szempontjából szükséges kritériumként fogalmazzák meg a bemeneti tesztek adott szint fölötti teljesítését. Előfordul, hogy ezekhez a tesztekhez mindenki által hozzáférhető gyakorló feladatokat is rendelkezésre bocsátanak vagy elérhetővé teszik a korábbi évek feladatsorait. Tekintettel arra, hogy a hazánkban tematikus és szintek szerinti bontásban az érettségizők rendelkezésére állnak a korábbi évek feladatsorai (`https://www.studiumgenerale.hu/matek-erettsegik/`), a tanulók rutinosak az ún. rátanulásban, amikor adott feladattípusok megoldásmenetét sajátítják el, nem egyszer mélyebb megértés nélkül. Több helyen bevett gyakorlat a mintadolgozat intézménye, ami esetleg azt is jelentheti, hogy a számonkérés alkalmával a diáknak egy a kiadotthoz teljesen hasonló feladatsorral kell megküzdenie. Ilyen mérések alkalmával azonban nem kapunk képet arról, miként fog teljesíteni a diák akkor, ha nemcsak begyakorolt rutinokból kell számot adnia a tudásáról. Éppen ezért ezek a mérések nem jelzik előre, hogy a hallgató meg tud-e birkózni a kalkulus tárgy nehézségeivel, sikeresen fogja-e venni azokat az akadályokat, amelyeket a matematikaigényes szaktárgyak jelentenek.

A Budapesti Műszaki és Gazdaságtudományi Egyetemen több éve foglalkozunk azzal, hogy az ütemterv szerinti haladást minél inkább előre jelző feladatsort készítsünk és ezzel mérjük az elsőéveseket. Az elmúlt években a mérőtesztek struktúrája is változott: az első idők matematika tesztjei nyelvi feladatokkal is kiegészültek. A pandémia okán nem minden esetben került sor jelenléti tesztelésre. Mostanra kellően nagy adatbázis áll rendelkezésünkre, amelyen nemcsak azt vizsgálhatjuk, milyen mértékben állnak készen a hallgatók arra, hogy sikerrel teljesítsék a kalkulus tárgyakat, hanem azt is, mennyire megbízható az online tesztelés, illetve milyen hatással volt a COVID-19 járvány az elsőévesek felkészültségére. Dolgozatunk célja, hogy modern adattudományi eszközök felhasználásával megvizsgáljuk a járvány kirobbanása előtt és után kezdett hallgatók bemeneti adataiban fellelhető változásokat, és ezen különbségek hatását az első félév végi teljesítményükre.

2. Irodalmi áttekintés

Nemzetközi téren a járvány diákokra gyakorolt hatásáról számos kutatás készült az elmúlt években. A Cavannaugh et al. [13] által vizsgált 837 amerikai tanuló hallgató esetén az átlagos tanulmányi átlag 0.1-gyel nőtt a pandémia alatt, és a növekedést legjobban befolyásoló tényező az egyes diákoknál az egyetemükön tanuló hallgatók száma volt. Ugyanakkor Alawamleh et al. [18] munkája, valamint Yamini Chandra [16] kutatása arra világítanak rá, hogy az online oktatás alatt gyakori volt a hallgatóknál a motivációhiány és az elidegenedés érzete is, valamint drasztikusan megnőtt a tanulókra és szüleikre nehezedő stressz illetve munka mértéke is. Onyema et al. [15] munkája azt a következtetést vonja le, hogy a digitális időszak negatívan érintette az oktatási tevékenységek jelentős részét, amely a tanulmányi teljesítmény csökkenése mellett diákhitel-növekedésben, egyetemi költségvetés-megvonásban és kutatási korlátozásokban is megnyilvánult. Ezek mellett az online oktatás hatékonyságát nehezítő körülmények közé sorolja a kiegyenlítetlen elektronikai infrastruktúrát, az oktatók elmaradott digitális készségeit, hálózati és szoftver elérhetőségi problémákat és más egyéb tényezőket is. Összességében elmondható, hogy járvány rávilágított a hagyományos oktatási rendszerek sebezhetőségére, ugyanakkor motivációt is adott azok átértékelésére, korszerűsítésére és új online oktatási módszerek kialakítására egyaránt [17, 19].

A pandémia hatásáról a diákok teljesítményére vonatkozóan hazai szinten is készültek kimutatások. Monostori Judit [26] jelentésében többek között az olvasható, hogy a 2020/2021-as tanévben a magyar alsó- és középoktatásban tanuló diákok teljesítménye lényegesen elmarad a jelenléti oktatás során megszokott szinttől, illetve a legnagyobb kihívást a tanulók figyelmének fenntartása és a szociális elszigetelődés leküzdése jelenti. Az eredmények alakulásában továbbá sokszor nagyobb jelentőséggel bír a szülőktől kapott otthoni segítség, mint az elektronikus infrastruktúra helyzete. Proháczik Ágnes [27] digitális oktatásról szóló felmérése arról számol be, hogy a diákok, szülők és tanárok túlnyomó része jelentős többletidő-ráfordítással tud csak megbirkózni az online tanulás nehézségeivel. A plusz ráfordított idő mennyisége régiónként eltérő, azonban szoros összefüggésben van mind a diákok, mind a tanárok digitális kompetenciájával, amelynek fejlesztése szerinte kulcsfontosságú. Csépe et al. [28] kutatásában a COVID-19 járvány hatását vizsgálták az egyetemi hallgatók életvitelére vonatkozólag, ahol azt állapították meg, hogy nőtt a nem dohányzók száma, a megkérdezettek közel 20%-a egészségesebb étrendre tért át, valamint a távolról bejáró hallgatók kiegyensúlyozottabb és egészségesebb alvási szokásokat tudtak felvenni.

A pandémia okozta feladatok és problémák megoldásában a gépi tanulási eljárások is fontos szerepet játszottak, legyen szó magáról a járvány terjedésének előrejelzéséről [23–25], vagy az álhírek és tévinformációk beazonosításáról [20–22]. Természetesen a problémák közé tartozik a diákok/hallgatók teljesítményének és lelki állapotának előrejelzése is, amelynek modellezésére több megoldás is született a járvány kezdete óta.

Niyogisubizo et al. [2] egy komplex, kétrétegű, hagyományos és mély tanulásos elemeket

kombináló osztályozó segítségével értek el 90% fölötti pontosságot a Nyitrai Konstantin Filozófus Egyetem hallgatóinak lemorzsolódását vizsgálva. Ugyanakkor a modellezésben olyan adatokat is felhasználtak, amelyekről csak a félév közepén kapunk információt. Tarik et al. [4] kutatása során hagyományos modelleket (mint döntési fák és lineáris regresszió) alkalmaztak marokkói hallgatók érettségi átlagának előrejelzéséhez, amely változó hatásfokot mutatott. A predikcióra nézve az egyes változók fontosságát is mérve Segura et al. [3] szintén többnyire hagyományos gépi tanulási algoritmusok segítségével értek el 90% fölötti eredményeket a spanyol 'Universidad Complutense de Madrid' egyetem hallgatóit vizsgálva lemorzsolódás terén, ahol a kutatás során 28-féle bemeneti változóval és különböző szakirányon tanuló hallgatók adataival dolgoztak. Dake et al. [5] Naive Bayes, véletlen erdő és döntési fa alapú algoritmusokat használtak a teljesítmény előrejelzésére. A kutatás során továbbá megállapították azt is, hogy a pontosság mint mérőszám nem feltétlen elegendő egy osztályozó jóságának eldöntésére. Atlam et al. [14] vizsgálataik során arab országokban tanuló hallgatóktól gyűjtött, a járvány megélésével és online oktatással kapcsolatos adatokat elemeztek különféle gépi tanulási algoritmusokkal és leíró statisztikai eszközökkel. A gépi tanulás során az egyes kérdésekre adott válaszokból a hallgatók mentális állapotának mint kategorikus változónak a prediktálása volt a cél, ahol a legjobb osztályozók 100%-os teljesítményt is képesek voltak elérni.

3. Adatprepació

Az elemzésünkben szereplő bemeneti változók két csoportba oszthatók. Egyik csoportba soroljuk azokat, amelyekkel a diákok már a felvétel pillanatában rendelkeznek, a másikba azokat, amelyek a belépést követően jönnek létre. Az először említett családba tartozik a tanulmányi pontszám, amely adott középiskolai tantárgyak több tanévbeli osztályzatainak összegéből képződik és az általunk vizsgált populációban maximálisan 200 pont lehet, az érettségi eredmény, amely a tantárgyakra kapott jegyekből számítódik és legfeljebb 200 pont lehet, valamint a többletpontok, amelyeket tanulmányi eredményért (emelt érettségi, versenyeredmények) vagy sport- és egyéb területeken elért kiemelkedő teljesítményért, nyelvvizsgáért lehet kapni összesen legfeljebb 100 pont erejéig. Ugyanezen család eleme továbbá, hogy milyen típusú osztályba járt a diák (matematika vagy természettudományi tagozatos, nem tagozatos), hogy emelt vagy középszinten érettségizett-e matematikából, illetve, hogy milyen szakra jár a VBK-n belül.

A belépést követően létrejövő változók családjában két elem szerepel: a 0. zárthelyi dolgozat (továbbiakban 0. ZH) eredménye és a matematikai-nyelvi teszten elért eredmény. Az előbbi teszt egy 45 perces, a középszintű érettségi anyagát számonkérő dolgozat, amely 15 tesztkérdésből áll, ahol minden feladatnál 5 lehetséges válasz közül kell kiválasztani az egyetlen helyeset. Az értékelésnél a hibás válaszokért -1, a helyes válaszért 4 pont jár. Ez egy ún. kritériumdolgozat az általunk vizsgált kar (BME Vegyészmérnöki és Biomérnöki Kar) hallgatói számára. Ez azt jelenti, hogy az első szemeszterbeli kalkulus (Matematika A1a) kurzusnál a vizsgára bocsátás szükséges feltétele a legalább 40%-os teljesítmény. A dolgozat megíratása papír alapon történik, továbbá minden elsőéves ugyanazt a 0. ZH-t írja meg, függetlenül az érettségi szintjétől és attól, hogy járt-e fakultációra.

A utóbbi teszt egy 80 perces, két nagy egységből (nyelvi és matematikai) álló, a magyar diákok számára szokatlan feladatokat is tartalmazó teszt. A magyar nyelvi feladatsor – hasonlóan a matematikához - nem csak a nyelvi tudást méri, hanem próbára teszi a különböző gondolkodási mechanizmusokat is. A teszt tizenhárom matematikai feleletválasztós kérdést tartalmaz, minden kérdés esetében 4 lehetséges válasszal, amelyek közül egy helyes van. A teszt három részből áll: az első blokk, amely az első négy feladatot foglalja magába az alapvető, procedurális számítási ismereteket ellenőrizi. Ezen feladatok hibátlan, esetleg egy hibával történő megoldása elvárható egy olyan személytől, aki műszaki képzésben kíván egyetemi tanulmányokat folytatni. A második blokk az első blokkhoz képest nehezebb feladatokat tartalmaz. Célja a Matematika A1 kurzushoz szükséges ismeretek meglétének ellenőrzése. Az első blokk feladataihoz képest ebben voltak összetettebb példák, de minden feladat megfogalmazása olyan volt, amellyel már középiskolában is találkozhatott a tesztet író. Ebbe a blokkba 6 feladat tartozik. Az utolsó, harmadik egység olyan szokatlan feladatokat tartalmaz, amelyeket csak a középiskolából származó alapos ismeretekkel rendelkező tanulók tudtak megoldani és azok, akik tudják alkalmazni ezt a tudást miközben váratlan feladatokat oldanak meg. Ezen feladatok megoldásához a legmagasabb szintű absztrakcióra volt szükség. Ez a blokk 4 feladatot tartalmaz. A matematikai-

nyelvi teszt matematikai része különböző volt aszerint, hogy a hallgató járt, vagy nem járt fakultációra, hogy mindenkinél a tőle elvárható szint meglétét ellenőrizzük. A nyelvi teszt feladatai viszonylag széles procedurális készségspektrumot fognak be. A nyelvtani elemek ismeretén túl a nyelvi absztrakció több területét érintik. A matematikai-nyelvi teszt elektronikusan, az EduBase online oktatási platformon került kitöltésre.

A bemeneti adatokat az EduBase és Neptun rendszerekből lekérve kaptuk több adattábla formájában. Az ezen adatokból nyert, később használt változók egy része alapvetően rendelkezésünkre állt a táblákból, másokat a meglévő sorokból és oszlopokból megfelelő *feature engineering* segítségével hoztunk létre. A változók modellezésben használt megnevezése és jellege az 1. táblázatban látható.

Emelt érettségi	Bináris változó arra vonatkozóan, hogy a hallgató matematikából emelt érettségit tett-e.
Matematika tagozat	Bináris változó arra vonatkozóan, hogy a hallgató matematika és/vagy természettudomány tagozatos volt-e.
Szak	Kategorikus változó, a hallgató szaka a VBK-n.
Matematika 1. blokk	Az elsőéves VBK hallgatók által írt kognitív teszt matematika részén az 1-4. kérdésekre adott helyes válaszok száma.
Matematika 2. blokk	A kognitív teszt matematika részén az 5-10. kérdésekre adott helyes válaszok száma.
Matematika 3. blokk	A kognitív teszt matematika részén a 11-14. kérdésekre adott helyes válaszok száma.
Kognitív eredmény	A matematika-nyelvi teszt nyelvi készségeket mérő részén elért százalékos teljesítmény (0-100-as skálán).
0. ZH pontszám	A BME központi 0. zárthelyi dolgozatán elért pontszám.
Tanulmányi pont	A felvételi pontszám tanulmányi pontokból származó része.
Érettségi pont	A felvételi pontszám érettségi pontokból származó része.
Többletpont	A felvételi pontszám többletpontokból származó része.
Matematika A1a	A Matematika A1a tárgyból szerzett érdemjegy.
Első féléves átlag	Az első félév végén megállapított kumulált átlag.
Kumulált átlag	A kutatás idejében aktuális félévig megállapított kumulált átlag.

1. táblázat. A vizsgált változók a két évben

A felderítő és modellezési fázisban használt adattáblát a rendelkezésre álló táblák megfelelő mértékű tisztításából, majd ezt követő Neptun-kód alapú összeillesztéséből nyertük. Ezen műveleteket Python-ban valamint R-ben végeztük el.

A kezdeti adattáblák közül a matematika és nyelvi eredményeket tartalmazó adatsor

tisztítására volt a legnagyobb szükség, ugyanis az EduBase rendszerben az egyes oszlopokra vonatozó mezőket a hallgatók töltötték ki, így a kategorikus változók nem voltak rendszerezve. Először egységesítettük a szakmegnevezést ("Vegyészmérnöki", "Biomérnöki", "Környezetmérnöki"), ahol pedig nem volt egyértelmű a szak, azt "UNKNOWN"-ra állítottuk. Ezt követően a teszten elért matematika és kognitív eredményt százalékos formátumról 0-1 közötti tizedestört formátumra változtattuk, valamint létrehoztunk 3 új oszlopot, amely a matematikateszt 3 blokkjában helyesen megválaszolt kérdések számát mutatja. Végül a vizsgálataink szempontjából irreleváns oszlopokat eltávolítottuk, valamint azon oszlopokat is kiszűrtük, amelyek értéke a többiből egyértelműen következtethető volt.

A kumulált átlagokat, felvételi pontszámokat illetve 0. ZH eredményeket tartalmazó tábláknál egy-egy hiányos és/vagy anomáliás sor, illetve az irreleváns oszlopok eltávolításán kívül nem volt szükség adattisztításra, a matematika jegyeket tartalmazó tábla pedig minden hallgató minden Matematika A1a tárgyból tett vizsgaalkalmáról és az azon szerzett érdemjegyről tartalmazott adatrekordot, így ezekből meg kellett határozni azt a végső érdemjegyet, amellyel a hallgató a tárgyat elvégezte. A végső összeillesztés során 10-20 fős sorvesztéssel is kellett számolnunk mindkét évben, ugyanis voltak olyan hallgatók, akikről nem minden táblában volt adat (vagy azért, mert nem írtak kognitív tesztet, vagy azért, mert az első félév vége előtt elhagyták a szakot). Az így kapott adattáblák, amelyek a 2019-es és 2021-es évet reprezentálják, rendre 230 és 202 adatrekordot tartalmaztak, melyekben a vegyészmérnök, biomérnök és környezetmérnök hallgatók száma rendre 119, 81, 28 illetve 104, 71, 21 volt.

A számszerű adatokat legtöbb esetben skálázásnak is alávetettük attribútumonként, amelyhez modelltől és feladattól függően kétféle módszert alkalmaztunk. Az egyik használt módszer a *min-max* skálázás volt, amely az adatokat a [0,1] intervallumra képezi le az alábbi módon:

$$x_i^{'} = \frac{x_i - min_{i\in\{1,\dots,n\}}(x_i)}{max_{i\in\{1,\dots,n\}}(x_i) - min_{i\in\{1,\dots,n\}}(x_i)} \tag{1}$$

ahol az $x_i^{'}$ az i-edik rekord, x_i értékének a képe. A másik módszer a *kvantilis alapú [0,1] uniform skálázás* volt, amelynél az egyes attribútumok értékeit az általuk meghatározott tapasztalati eloszlásfüggvénybe helyettesítjük be, így az adatok egy [0,1] intervallumon értelmezett uniform eloszlás realizációi lesznek.

4. Felderítő adatelemzés a két évben

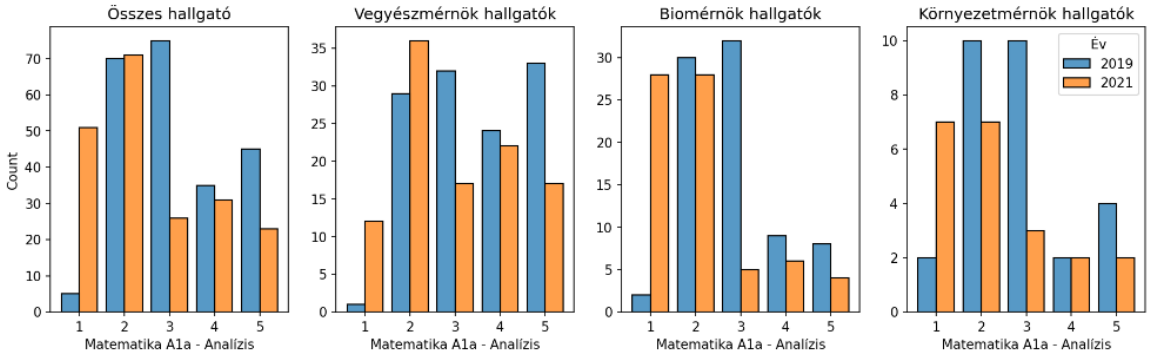
Egy részben feltáró adatelemzés fontos szakasza a felderítő elemzés. Célja az adatok ábrázolása egyszerűen és gyorsan annak érdekében, hogy az adattípusok egyszerű viszonyai szemléletesen áttekinthetőek legyenek. Erre a célra alkalmasak az oszlopdiagramok, szórásdiagramok, folyamatábrák és más grafikonok, továbbá klaszterezéssel kevésbé triviális összefüggések is kinyerhetők az adatból.

4.1. Általános ábrák

Ebben az alfejezetben tekintjük át a számos készült ábrából a leginformatívabbakat. Viszonylag sok változóval kellett dolgozni, így célszerű volt több részábrás egységekbe gyűjteni a hasonló felépítésű grafikonokat. A megfelelő ábrák elkészítéséhez a Python *matplotlib* és *seaborn* csomagjait használtuk. A változók értékeinek eloszlását oszlopdiagramokkal, a változópárok egymás közötti viszonyait szórásdiagramokkal szemléltettük.

Az 1. táblázatban felsorolt változók közül egyesekből aggregáltunk további két változót. A *Tanulmányi-* és *Érettségi pontot*, illetve a *Többletpontot* összeadva határoztuk meg a *felvételi összpontszámot*. Továbbá a *Matematika 1., 2.* és *3. blokk* értékeit összeadva kaptuk a *matematika összpontot*. Végül használtuk a *teljes teszteredményt*, ami rendelkezésre állt az eredeti adattáblákban, de a *kognitív eredmény* és a *Matematika 1., 2.* és *3. blokk* lineáris kombinációjaként is számolható lett volna.

Az 1. ábrán a két évben szerzett Matematika A1a -Analízis jegyek eloszlása látható összesítve és szakokra lebontva a két évben.

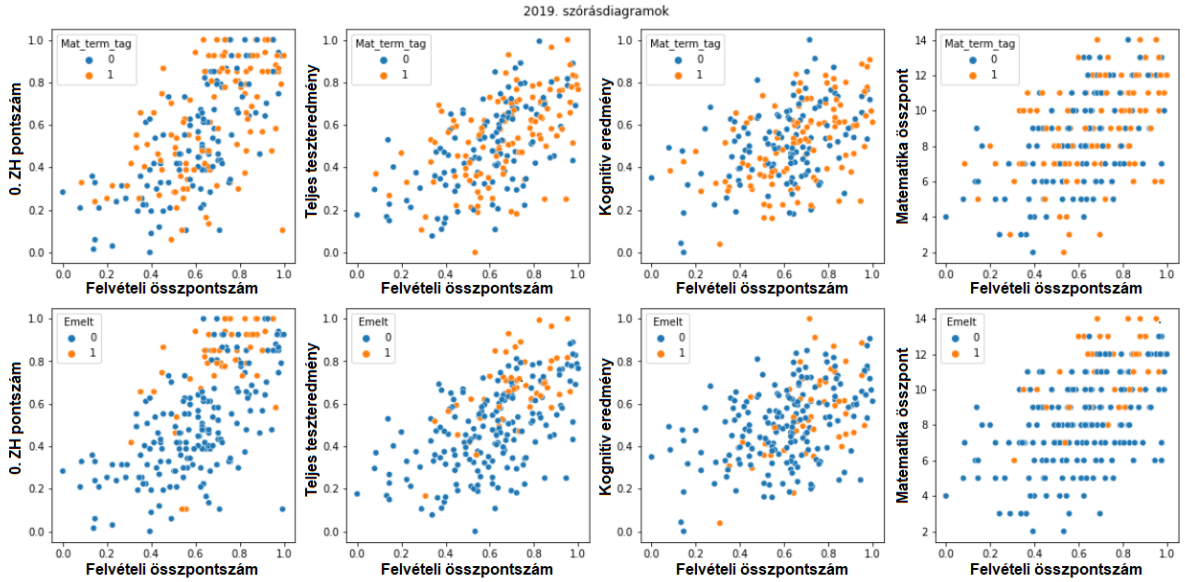


1. ábra. Az elsőéves matematikajegyek eloszlása az egyes években

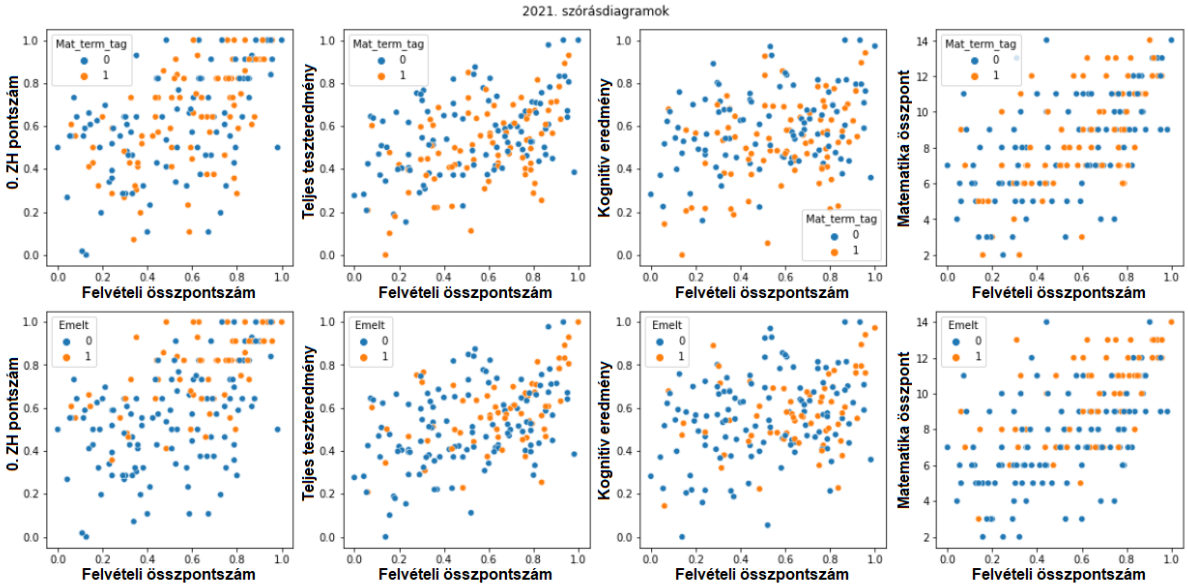
A legszembetűnőbb különbség a két év között az elégtelen, elégséges és közepes érdemjegyek eloszlásának tekintetében van. Az összesített ábrán jól látható, hogy míg 2019-ben a 2-es és 3-as érdemjegyek vannak túlsúlyban, addig 2021-ben az 1-es és 2-es érdemjegyek kerültek többségbe. Ugyanez a jelenség figyelhető meg a biomérnök és környezetmérnök hallgatók esetén is.

A 2. és 3. ábrán a 0. ZH, a teszt összpontszáma és a két külön rész pontszámai vannak ábrázolva pontfelhőként a felvételi összpontszám függvényében a két évben, min-

den esetben kétszer. A felső sorban a pontok az alapján vannak színezve, hogy járt-e természettudományi tagozatra a hallgató (1 igen, 0 nem), az alsóban pedig, hogy emelt szinten érettségizett-e (1 igen, 0 nem).



2. ábra. 2019-es adatok szórásdiagramjai

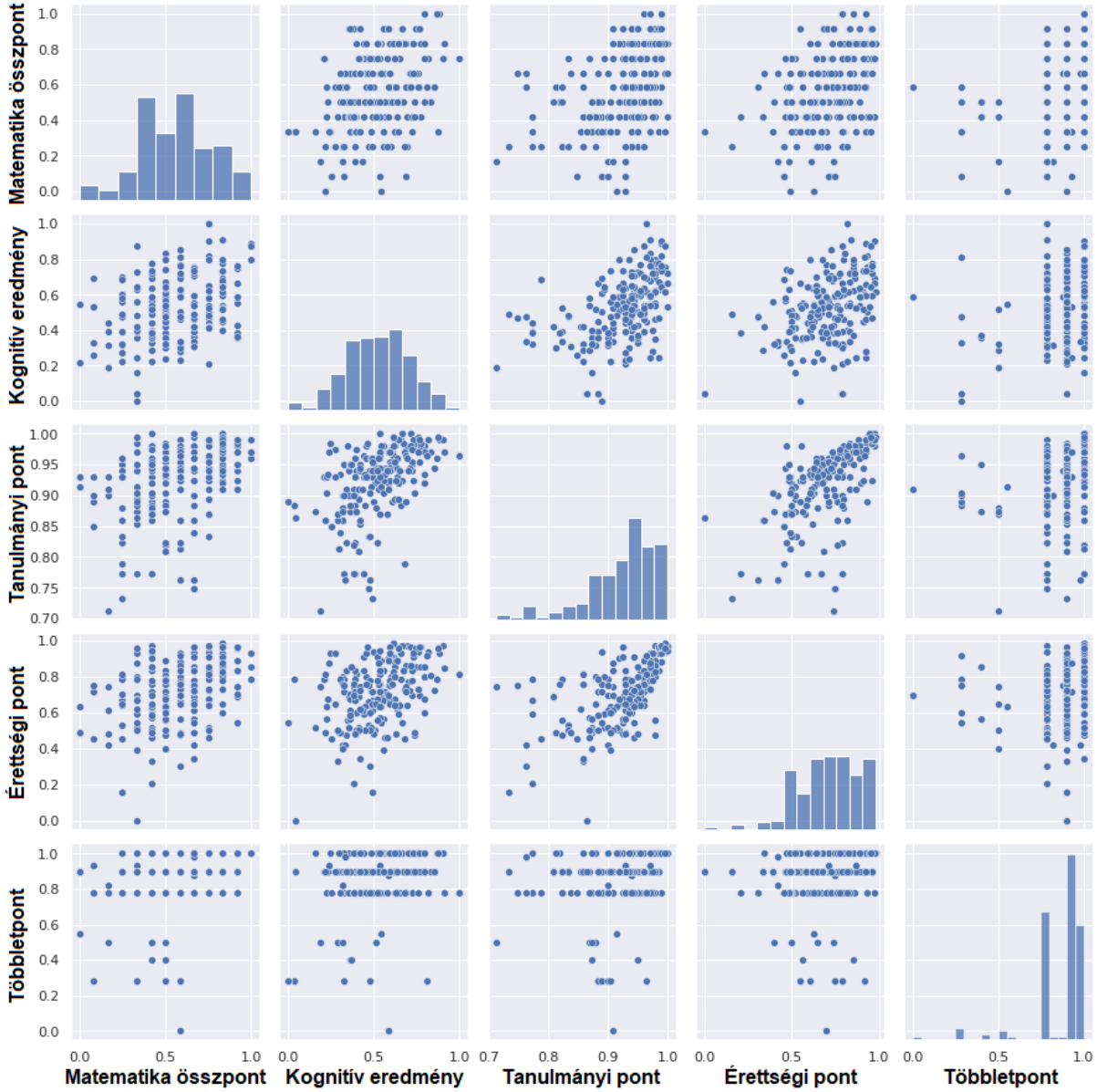


3. ábra. 2021-es adatok szórásdiagramjai

A 2019-es ábrán egyrészt az látszik, hogy minden részábrán a bal felső és a jobb alsó sarkokban többnyire nincs pont, ez a pozitív korreláció jele. Másrészt a színezésből arra következtethetünk, hogy az emelt érettségi megléte (amelyért ráadásul 50 többletpont jár, ha a diák 40%-nál jobb eredményt ér el) és a magas felvételi pontszám további kimagasló eredményekkel társul, kivéve a teszt nyelvi részénél, ott jobban megoszlanak az eredmények. Ezzel szemben a természettudományi tagozatoknak nincs látható hatása a további

eredményekre, ugyanis a sárga pontok szét vannak szóródva az egész pontfelhőre minden esetben.

A 2021-es ábráról az állapítható meg, hogy a 2019-es ábrához képest nőtt a változók szórása, ugyanis a pontfelhők jobban szétterülnek. Ennek a hátterében állhatnak a vírushelyzet okozta nehézségek és az adott években bevezetett távolléti oktatás. A természettudományi tagozaton való tanulás továbbra sem fejtett ki hatást a többi eredményre. Az emelt szintű érettségi meglétének hatása viszont 2019-hez képest csökkent, ez a teljes teszteredménynél a legszemléletesebb.

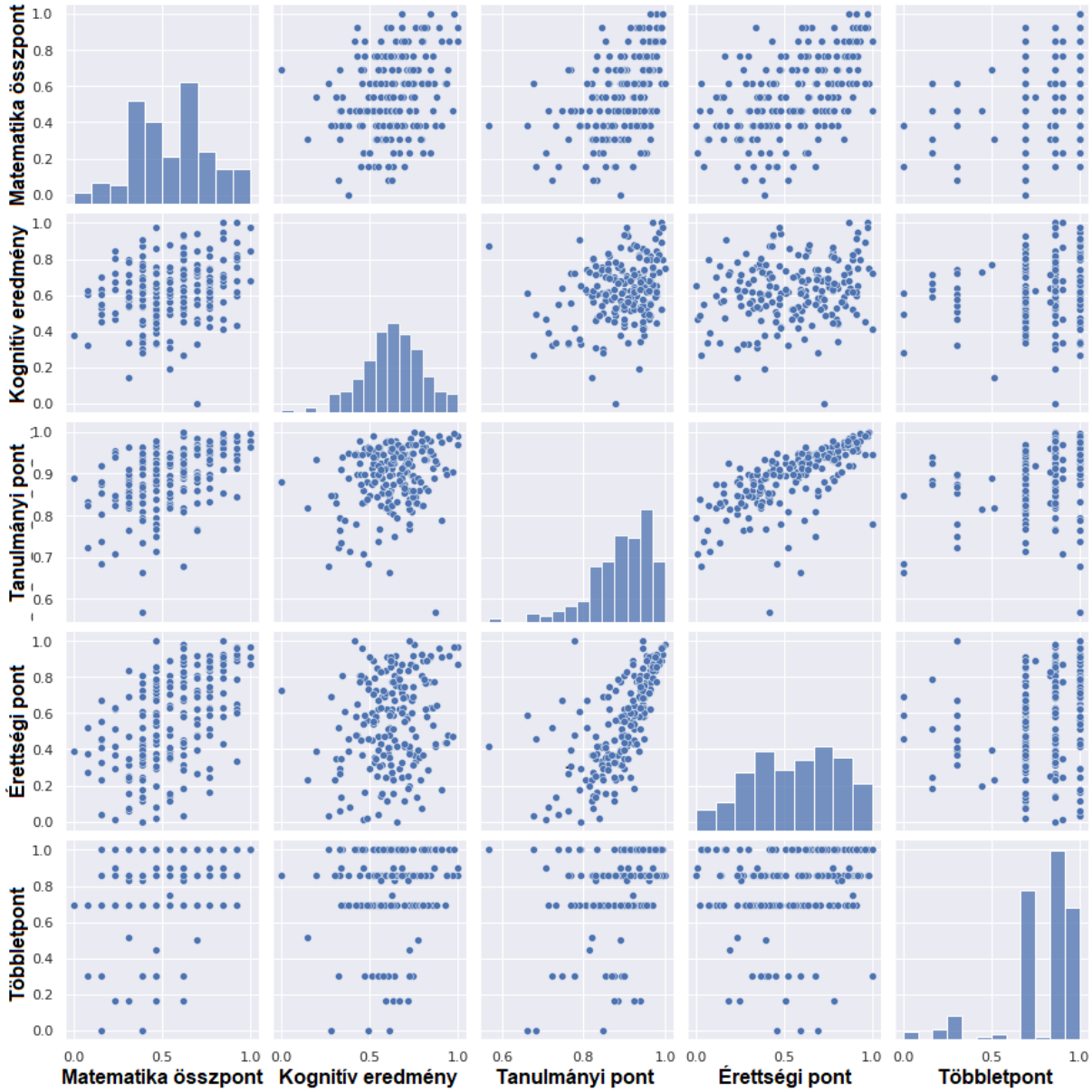


4. ábra. 2019. Teszt és felvételi pontszám részei

A 4. ábrán látható a *seaborn pairplot* függvény egyik eredménye. Ebben az esetben volt két hallgató, aki pont duplázással számította a teljes felvételi pontszámát, így a hozott pontszámánál 0 állt értékként. Ez természetesen torzítást okozott a megfelelő

részábrákon, így ki lettek véve. A változók itt min-max skálázva vannak, így minden tengelyen 0-tól 1-ig vannak értékek.

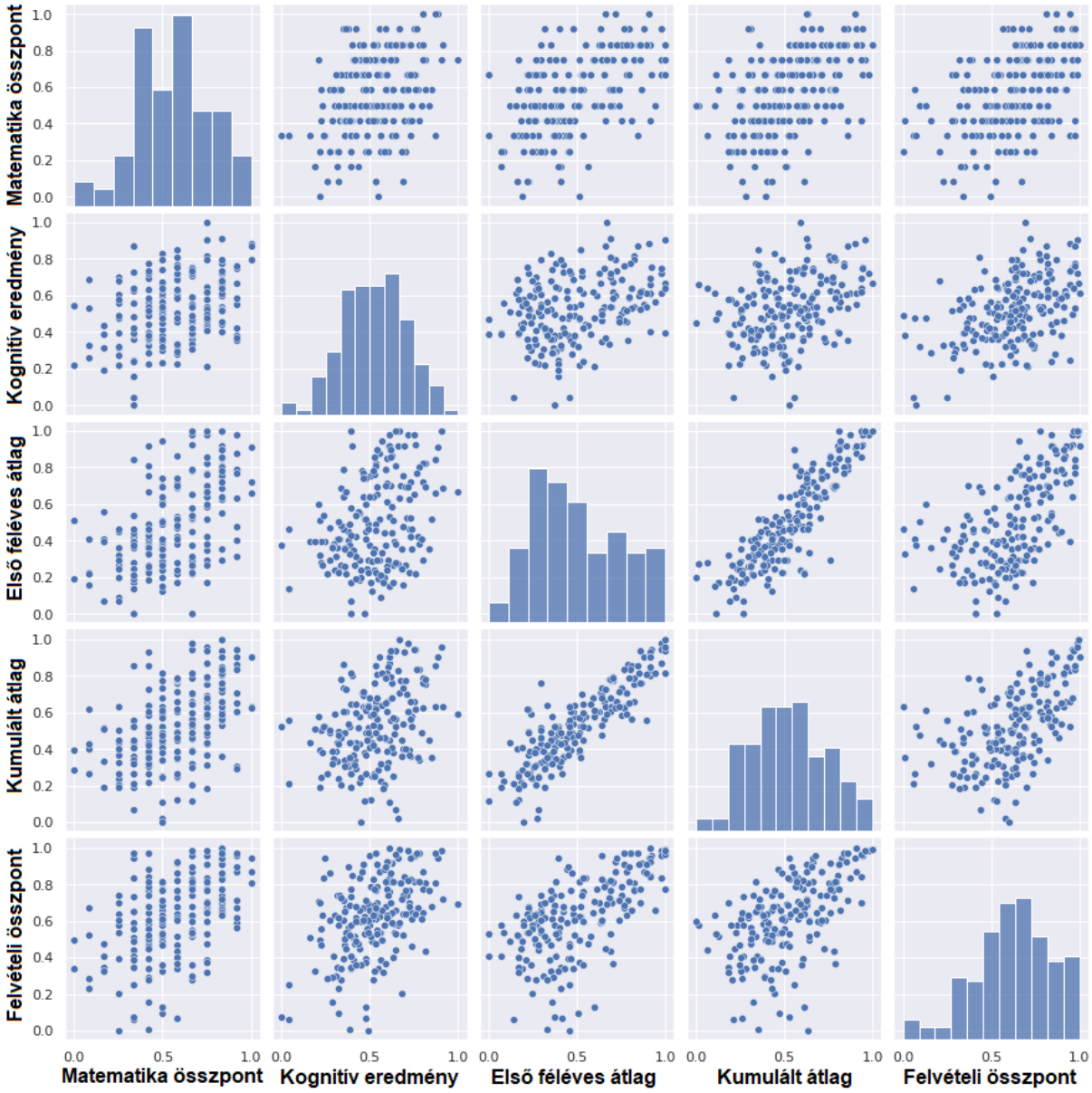
A többletpontot tartalmazó részábráktól és az oszlopdiagramoktól eltekintve, a többi részábrán a pontfelhők gyenge pozitív korrelációt mutatnak, de az adatok szórása minden dimenzióban egészen nagy. Kiemelendő, hogy a kognitív eredmény oszlopdiagramja a normális eloszlás haranggörbéjéhez közelít annak ellenére, hogy a felvételi pontszám részeinek eseteiben a jó eredmények dominálnak.



5. ábra. 2021. Teszt és felvételi pontszám részei

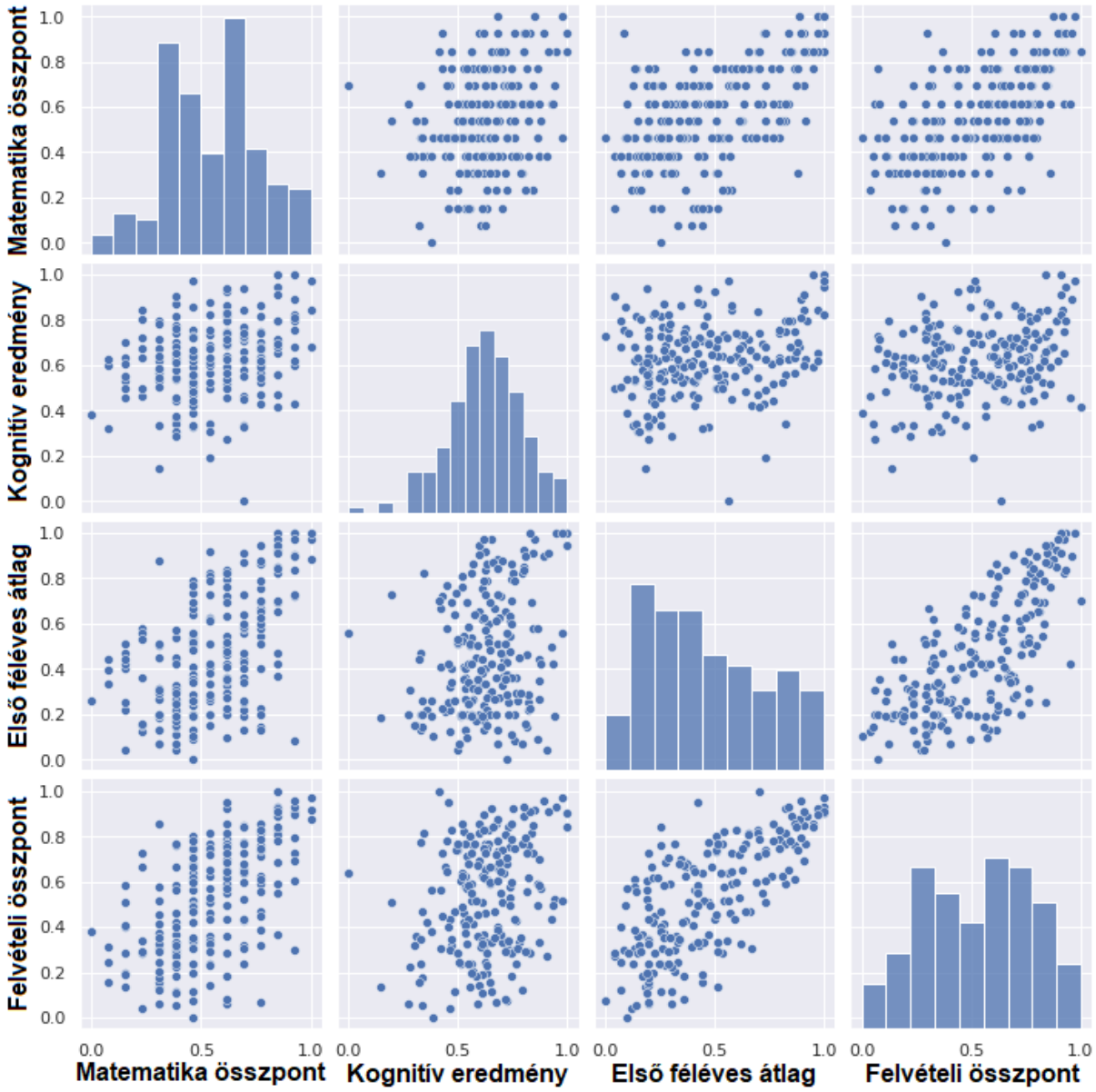
Az 5. ábrán egészen hasonló képet látunk, mint a 2019 esetében (4. ábra), de itt ismét erősebb a szórás. Továbbá szembetűnő az érettségi pontszámok eloszlásának megváltozása. Míg 2019-ben a skálázott adatok nagy része 0.5-1 között van, 2021-ben már az egész intervallumra széthúzódnak.

Eddig a pontfelhők nem mutatnak erős korrelációt, illetve a kétdimenziós vetületekben nem emelkednek ki csoportok. Más változókat is vizsgáltunk, a róluk készült ábrák alább tekinthetők meg.



6. ábra. 2019. Teszt részek, átlagok és felvételi összpontszám

A 6. ábrán a felvételi pontszám össze lett vonva és be lett véve az első féléves és a későbbi kumulált átlag. Nyilván az utóbbi kettő erősen összefügg. Enyhén kirajzolódik egy összefüggés a felvételi pontszám és az átlag között. Megjegyzendő továbbá, hogy a kumulált átlag eloszlása jobban közelít a normális eloszláshoz, mint az első féléves átlagé, ami eltolódik a skála alja felé.



7. ábra. 2021. Teszt részek, átlag és felvételi összpontszám

A 7. ábra esetében nem volt értelme a kumulált átlagot bevenni, ugyanis a kutatás idejében megegyezett az első féléves átlaggal. Itt is megmaradt a felvételi összpontszám és az átlag közötti összefüggés, de a pontfelhő kevésbé zajos. A matematikai rész pontszáma és az átlag között is van gyenge, pozitív összefüggés.

4.2. Folyamatábrák (Sankey-diagramok)

A Sankey-diagramok olyan folyamatábrák, amelyekben az ábrázolt folyamok szélessége arányos a megfelelő folyamértékekkel. Ezáltal könnyen ábrázolhatóak adott állapotok közötti folyamok, illetve ezek egymással vett aránya. Egy hátránya van, hogy egy ábrán csak két osztálycsoport között olvashatjuk le az áramlásokat. Ha egy harmadik osztálycsoportot is belehelyezünk, és csak a másodikkal kötjük össze, akkor az első és a harmadik

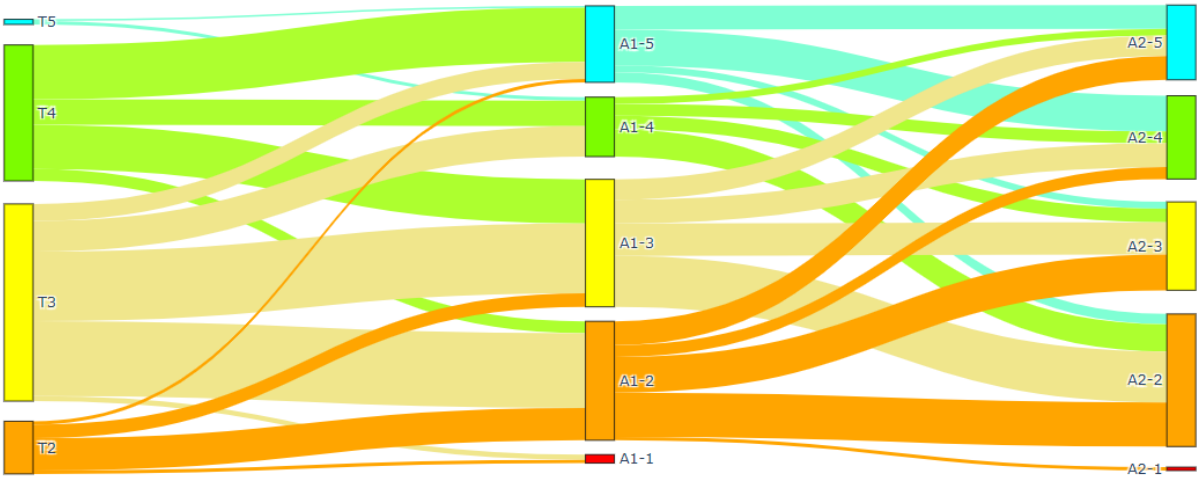
közötti vándorlások nem jelennek meg rajta. Emiatt van némi redundancia a lentebb ismertetett ábrákban.
Dolgozatunkban öt opciót vizsgáltunk a két évben az elérhető adatoknak megfelelően. Mindkét esetben az állapotok a hallgatók valamely eredményét jelentették, például felvételi összpontszámot.

- 2019:
    1. teljes teszteredmény - *A1* jegy - *A2* jegy.
- 2019 és 2021:
    1. tanulmányi pont - érettségi pont - *A1* jegy.
    2. felvételi összpontszám - 0. ZH - teljes teszteredmény.
    3. felvételi összpontszám - teljes teszteredmény - *A1* jegy - kumulált átlag.
    4. felvételi összpontszám - teljes teszteredmény - első féléves átlag - kumulált átlag.

Az első esetet csak folyamatábra erejéig vizsgáltuk, ugyanis csak a 2019-es évfolyamra volt meg az $A2$ jegy, hiszen a 2021-es évfolyam még nem vehette fel a tárgyat a kutatás idejében. Sőt, a 2021-es évfolyamnál megegyezett az első féléves és a kumulált átlag, így az utolsó esetben az utóbbi értelemszerűen nem lett még egyszer beletéve.

Az *A1*, *A2* jegyeken kívül a többi változó folytonos volt, így szükség volt ezek diszkretizálására. Öt-öt osztály lett létrehozva minden esetben, de nem feltétlenül került mindegyikbe hallgató. Az átlagok kerekítve lettek, a felvételi összpontszám, teljes teszteredmény és a 0. ZH pontszámok ekvidisztáns módon lettek felosztva.

A *felvételi összpontszám - 0. ZH - teljes teszteredmény* változatban a 0. ZH és a teszt eredményei nullától az elérhető pontszámig lettek felosztva 20%-os lépésközökkel. A felvételi összpontszám a legalacsonyabb értéktől 500 pontig. A többi változat a kutatás későbbi szakaszában készült, így a folytonos mutatók az értékkészletük terjedelme szerint lettek felosztva, nem az elérhető pontszámok alapján. Végül a kirajzolódott ábrákat kézileg még igazítani kellett, ugyanis az eredmények csoportjai nem feltétlenül álltak megfelelő sorrendben a függőleges tengelyen, illetve azokban az esetekben, amikor nem volt mind az öt osztályban hallgató, egy-egy csoport átcsúszott egy másik eredmény oszlopába. Ezek azonban nem okoztak gondot, ugyanis a csoport mozgatásával a hozzátartozó sávok is automatikusan együtt mozogtak.

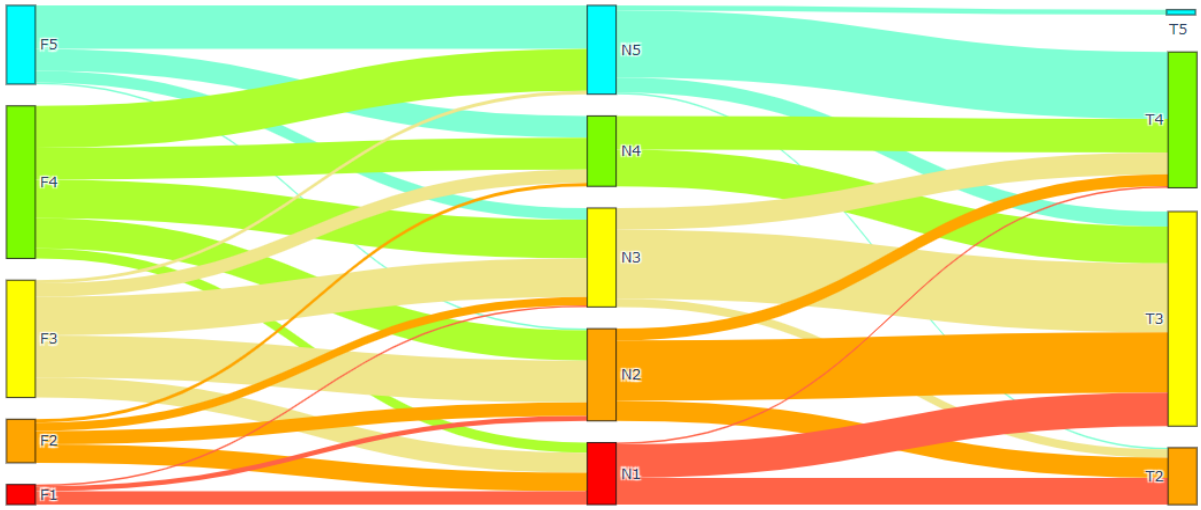


8. ábra. 2019. teljes teszteredmény - *A1* jegy - *A2* jegy

A 8. ábrán rögtön látható, hogy a kognitív tesztben mindenki 20% felett teljesített, de nagyon kevés hallgatónak sikerült 80%-nál magasabb eredményt elérni. Ehhez képest az tantárgyakon elért jegyek eloszlása egyenletesebb.

Látható, hogy akik a teszten gyengébben teljesítettek, azok a későbbiekben többnyire rosszabb jegyeket szereztek, de nem kevesen javítottak is. A jobban teljesítők szintén tartották általában a szintet, de itt sem elhanyagolható azok száma, aki rontott. Meglepő a kettesről négyes-ötösre javítók és az ötösről rosszabb jegyekre rontó hallgatók aránya a teljes viszonylatban.

Összességében az figyelhető meg, hogy míg az első két változó közötti átmenet viszonylag egységes, a két tárgy közötti átmenetek nagyon változatosak. Ez azt jelenti, hogy a bemeneti teljesítmény részben befolyásolja az első néhány eredményt, de a későbbiekbe már kevesebb beleszólása van.

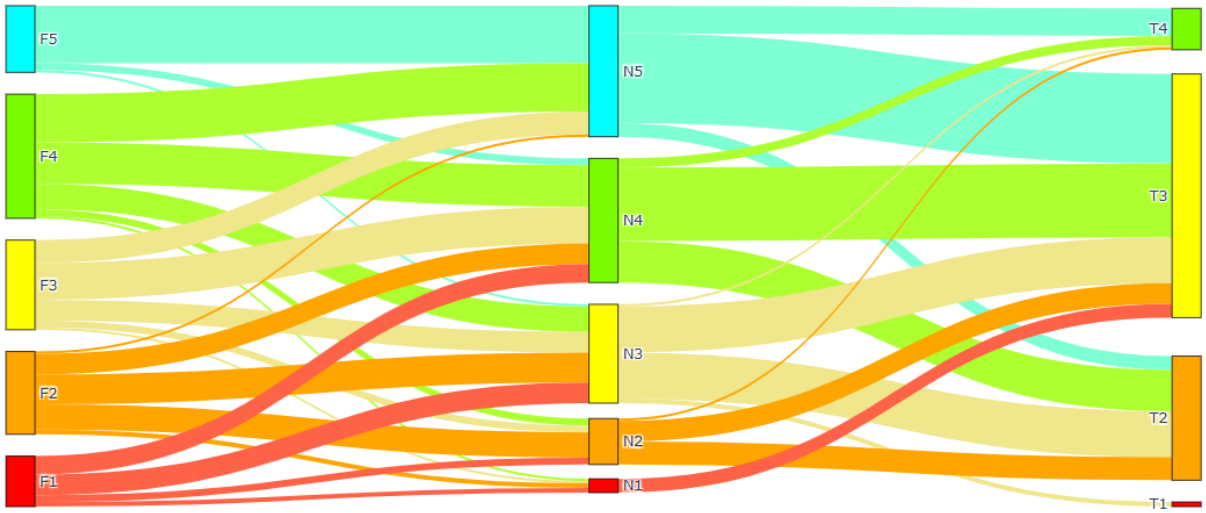


9. ábra. 2019. Felvételi összpontszám - 0. ZH - teljes teszteredmény

A 9. ábrán főleg a 0. ZH-ból a tesztbe menő folyamokon látszik egy erős tendencia.

Igaz itt még a régebbi, teljes szerezhető pontszám szerinti felosztás van, így csak négy osztály jött létre, de nagyon kevés többosztálynyi ugrás van, tehát feltehető, hogy erősen összefügg a 0. ZH és a teljes teszteredmény.

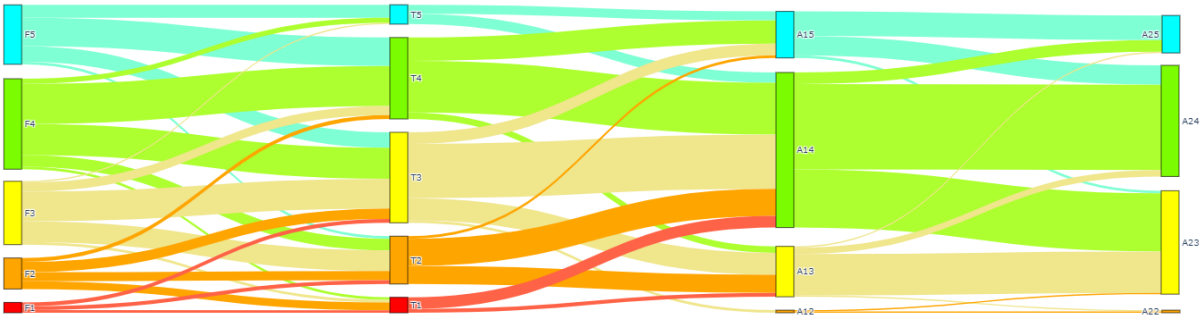
Ami a felvételi és 0. ZH pontszámok kapcsolatait illeti, a folyamok egészen szerteágazóak, de itt is fennáll, hogy a közel hasonló teljesítmények között szélesebbek, mint a nagyobb különbségek esetén.



10. ábra. 2021. Felvételi összpontszám - 0. ZH - teljes teszteredmény

Annak ellenére, hogy a tesztes oszlop a 2019-es tesztes oszlopra hasonlít, csak fejjel lefelé, ez a valós eloszlás. Valóban, míg 2019-ben nem volt 20% alatti eredmény, itt fordítva, nem volt 80% fölötti. A felvételi pontszámok arányai romlottak, arányosan nagyobbak lettek az alsó osztályok. Ellenkezőleg, a 0. ZH eredményei javultak, nagyobbak a magasabb pontszámhoz tartozó osztályok arányai.

Ami a folyamokat illeti, többnyire hasonlóak a tendenciák, bár fontos észrevétel, hogy akiknek a legrosszabb lett a 0. ZH eredményük, a kognitív teszt közepesen ment. Itt is elmondható, hogy a jobban teljesítő hallgatók többnyire mindenből jól teljesítettek, illetve a gyengébbek között ritkábbak voltak a jobb eredmények. Ennek ellenére itt is szerteágazóak a folyamok, tehát összefüggés van, de nem nagyon erős.

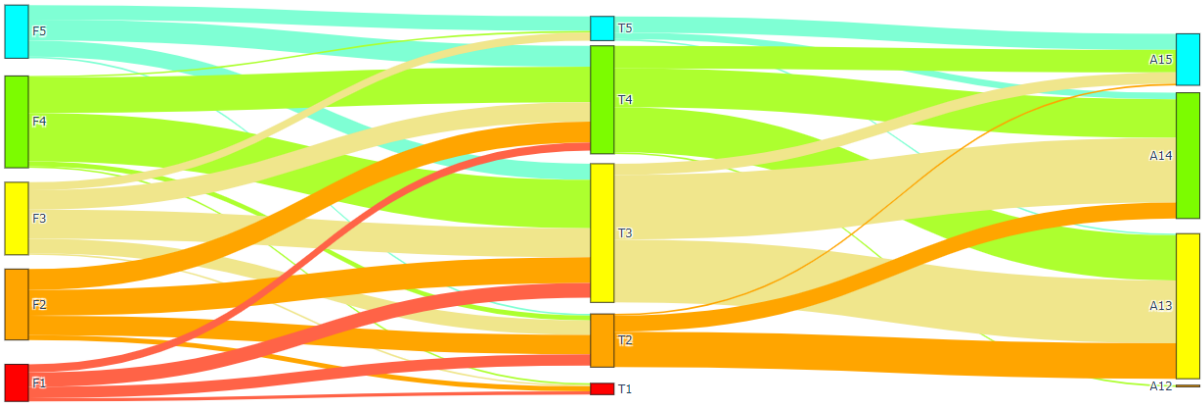


11. ábra. 2019. Felvételi összpontszám - teljes teszteredmény - első féléves átlag - kumulált átlag

A 11. ábrán már közvetlenül láthatóak a felvételi és kognitív teszt eredményeinek viszonyai, azonban itt már az értékkészlet van felosztva ekvidisztánsan. Ez abban is megnyilvánul, hogy a teszt pontszámosztályainak mások az arányai.

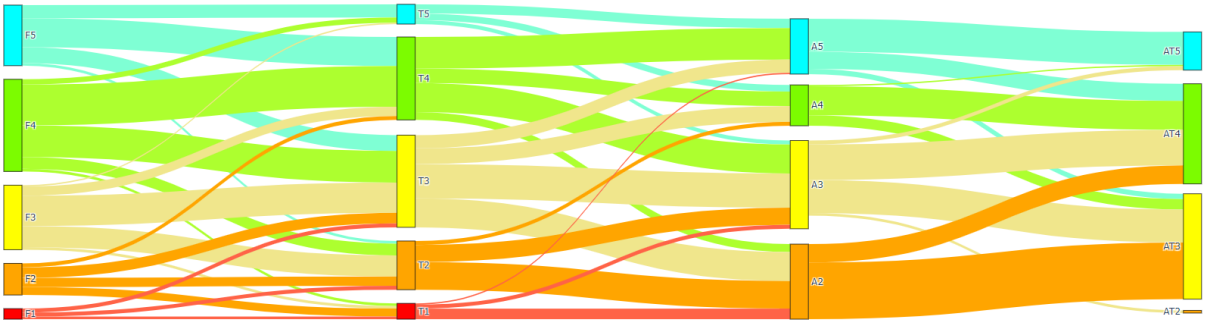
Két előnye volt az értékkészlet szerinti osztályozásnak: egyrészt gyorsabb volt (egy függvényre volt csak szükség), másrészt így az eredmények egymáshoz képest való relatív eloszlása jobban látszik. Míg a korábbi ábrákon nagyon nagy a hármas osztály, mert a legtöbben 40-60% körüli eredményeket értek el, itt az látszik, hogy a legjobb és legrosszabb dolgozatok pontszámai között, a terjedelem 20%-ával lépkedve milyen a hallgatók eloszlása. Így az lesz valóban "középtájban", aki a többi hallgató pontszámához képest átlaghoz közeli pontszámot szerzett, nem pedig az a sok tanuló, aki 40-60%-ot ért el.

Visszatérve a 11. ábrára, a fentebb megfigyelt tendenciák lényegében ismétlődnek, jó eredményeket jók, rosszakat rosszak követik nagyrészt, de továbbra is változatosak a folyamok. Összefüggőség szempontjából jó jel, hogy a legjobb teszteredményt elérők négyesnél rosszabb átlagot nem értek el. Sőt nagyon sok hallgatónak volt négyes az átlaga az első félévben, ami nagyrészt kétfelé oszlott az idő elteltével, többen megtartották, de sokan rontottak hármasra, páran javítottak. A többi átlagértékkel hasonló a helyzet, inkább megmaradt, mint javult vagy romlott.



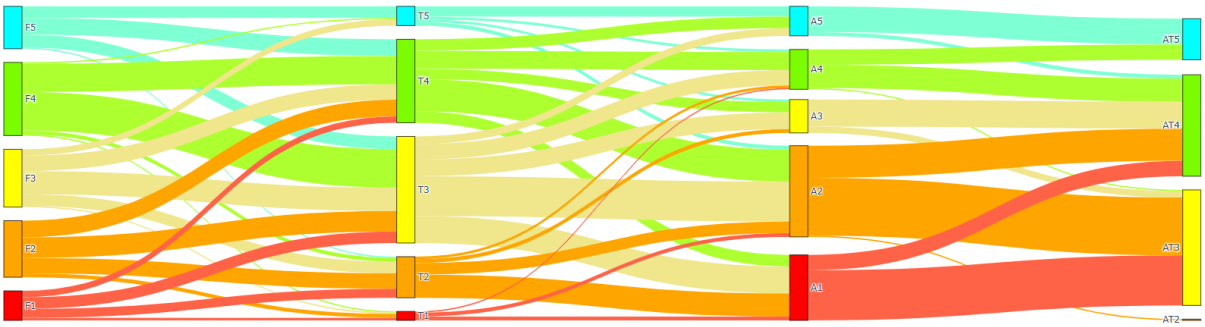
12. ábra. 2021. Felvételi összpontszám - teljes teszteredmény - első féléves átlag

A kutatás idejében ez az évfolyam még az első évében volt, így csak első féléves átlagokról volt adatunk. Ettől eltekintve, a 12. ábrán látszik, hogy 2019-hez képest több hallgatónak volt relatíve gyengébb a felvételi pontszáma, de ennek megfelelően többen is "javítottak". Az átlagok eloszlása romlott, aránylag több hármas és kevesebb négyes, illetve ötös lett.



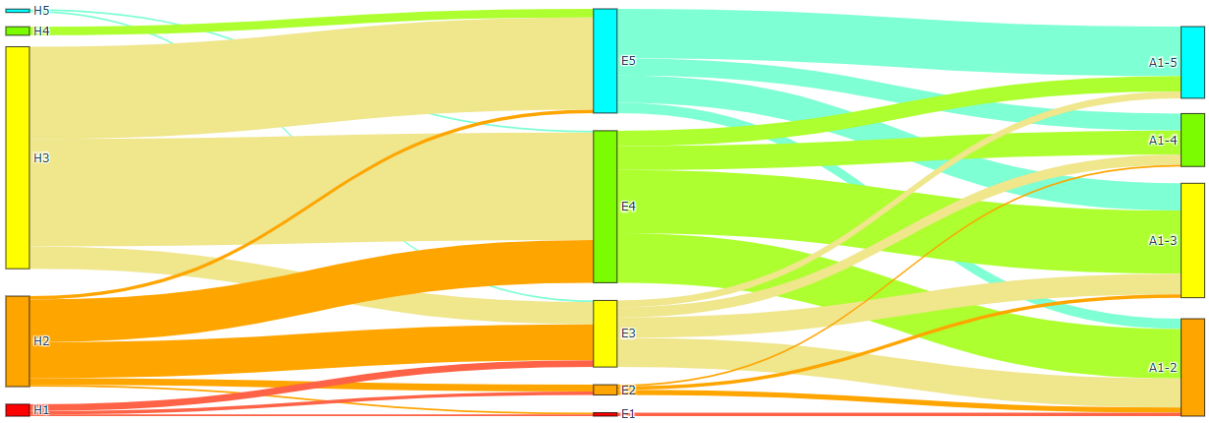
13. ábra. 2019. Felvételi összpontszám - teljes teszteredmény- *A1* jegy - kumulált átlag

A 13. ábra első két oszlopának viszonyát már láttuk, az újdonság az *A1* jegy az első féléves átlag helyett. Ettől függetlenül a tendenciák változatlanok, azok szereztek jobb, illetve rosszabb jegyet, akik rendszerint jobban, illetve rosszabban írták meg a tesztet. A trendet továbbra erősíti a jegyek és átlagok viszonya, bár értelemszerűen nem meglepő, hogy a két dolog összefügg.



14. ábra. 2021. Felvételi összpontszám - teljes teszteredmény - *A1* jegy - első féléves átlag

A 14. ábrán szembetűnő, hogy az *A1* jegyek valamilyen okból kifolyólag erősen romlottak. Míg 2019-ben nem bukott egy hallgató sem, itt majdnem a csoport negyedének nem sikerült elvégezni a tárgyat, sőt olyanoknak sem, akik a legjobbak között voltak a teszt alapján. Továbbá a csoportnak több, mint a fele nem tudott kettesnél jobb jegyet szerezni. Fentebb láttuk, hogy az átlagok is romlottak, de ezek szerint nem olyan mértékben, mint az *A1* jegyek. Kiemelendő, hogy az egyest és kettest szerzett hallgatók mégis nagyrészt hármas, sőt nem kevesen négyes átlagot szereztek. Ettől függetlenül a szokásos trendek itt is megmutatkoznak.

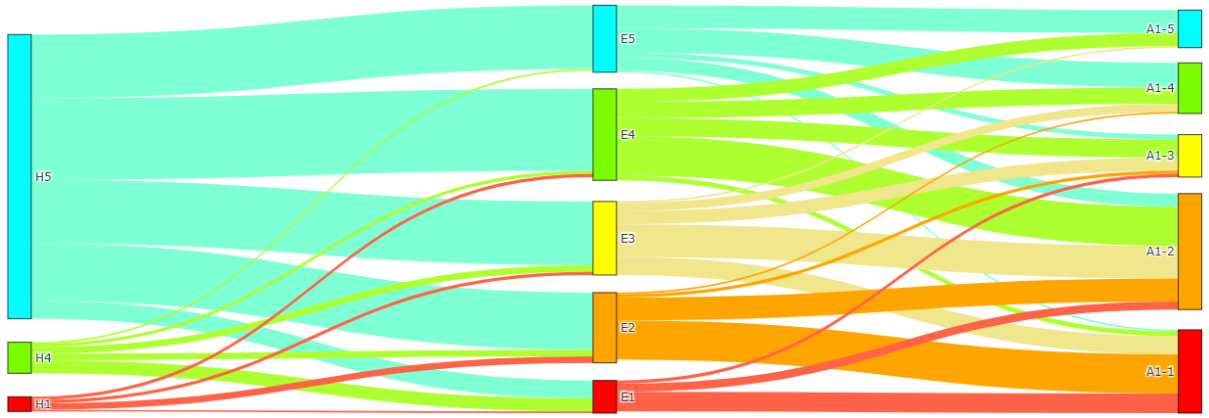


15. ábra. 2019. tanulmányi pont - érettségi pont - *A1* jegy

Ahogy később látni fogjuk, a modellek alapján a tanulmányi és érettségi pontszámok erősen befolyásolták a jegyekre vett predikciókat, így utólag elkészült a 15. és a 16. ábra.

Jellegzetes, hogy a hallgatók nagy részének hasonló volt a tanulmányi pontszáma, nagyon kevésnek volt kiemelkedő vagy relatíve kicsi. Aránylag kevés hallgatónak volt a legjobb pontszám 60%-a alatt az érettségije. Még ezek mellett is megőrződött az a tendencia, hogy egy-egy hallgató nagyrészt hasonlóan teljesít a különböző mutatók alapján.

Az ábráról leolvasható, hogy míg az utolsó oszlop felső felébe szinte csak kék és zöld folyamok érkeznek, mégis szerteágaznak, és például a legjobb érettségi pontszámokkal rendelkezők közül is többen kettest szereztek a tárgyból.



16. ábra. 2021. tanulmányi pont - érettségi pont - *A1* jegy

A 16. ábrán rögtön szembetűnik egy meglepő dolog, a tanulmányi pontszámoknál van egy óriási ötös osztály. Az értékeket leellenőrízve, tényleg nagyon sok esetben 180 és 190 között mozgott a hozott pontszám.

Ehhez (és a 15. ábrához) képest, az érettségi pontszám viszont egészen egyenletesen oszlott el. Az ábra bal felétől eltekintve, a teljesítmény megmaradására vonatkozó tendencia gyengébben, de ismét jelen van.

4.3. Klaszterezés

A klaszterezés célja általában esetleges nem triviális összefüggések feltárása, illetve az adatok csoportokba való sorolása gépi tanulási módszerekkel úgy, hogy az azonos csoportban lévő adatok "hasonlóbbak" legyenek egymás közt, mint a különböző csoportokból vett adatok.

Egyes esetekben a klaszterezés eredménye rámutathat néhány változóra, amelyeknek erős a magyarázóereje, így a további vizsgálatokból el lehet hagyni a többi változót. Ennek ellenére gyakoriak azok az esetek, amikor nehéz vagy nem lehet megállapítani ezeket a változókat, mert minden változó fontos. Ilyenkor az adatok két-három dimenziós vetületeinek szórásdiagramjai többnyire homogének és nem csoportosíthatóak szemre.

A 4.1. alfejezetben szereplő ábrákat áttekintve, első ránézésre nem rajzolódott ki olyan kétdimenziós vetület, amelyben láthatóan megjelennének csoportok. Ennek ellenére három gépi tanulási módszer lett kipróbálva:

1. K-közép algoritmus
2. DBSCAN algoritmus
3. Ward-féle hierarchikus klaszterező algoritmus.

Számos optimalizációs kísérlet volt: hiperparaméterek beállítása, változók válogatása, illetve többféle skálázás. Volt egy kísérlet PCA alkalmazására, azonban minden esetben csak az összes megadott változó reprezentálta elég jól az adatot, így ez el lett vetve. Sok modell és klaszterkép készült, a legérdekesebbek lentebb lesznek bemutatva. A szemléltetéshez a bevett változók közötti páros szórásdiagramok lettek legenerálva és a pontok az alapján lettek színezve, hogy mely klaszterbe kerültek.

4.3.1. K-közép algoritmus működése

Az algoritmus működési elve nagyon egyszerű. Az alapvető paraméter, amelyet meg kell adni, a létrehozandó osztályok kívánt száma, legyen ez k. Első lépésben választ a térben véletlenszerűen k pontot, ezek a klaszter középpontok, majd az összes pontot ahhoz a a középponthoz rendeli, amelyhez a legközelebb van. A második lépésben a kialakult osztályok középpontjait újraszámolja és azokkal a középpontokkal megismétli az első lépést. Ezt a két lépést iterálja addig, amíg két egymást követő osztályozás között már nem lesz jelentős változás.

Így alapvetően a legfontosabb feladat a megfelelő klaszterszám meghatározása. Ehhez létezik egy könyökszabály, melynek a lényege a következő: lefuttatjuk a klaszterezési algoritmust 1-től l-ig (l egy tetszőlegesen választott szám) megadva a klaszterszámokat, majd ábrázoljuk ezek esetében a pontoknak az átlagos négyzetes eltérését a saját klaszterközéppontjuktól. Természetesen ahogy több csoportra bontjuk a pontokat, úgy csökken ez az eltérés, így az a klaszterszám lesz optimális, ahol jelentős a javulás, de utána már nem. Ez

megfelelő esetben vizuálisan egy behajlított karra emlékeztet, innen ered a neve. Nehezen klaszterezhető adatok esetén ez a görbe "simább", nehezen állapítható meg a "könyök".

4.3.2. DBSCAN algoritmus működése

A DBSCAN (Density-based spatial clustering of applications with noise) sűrűség alapú klaszterező algoritmus [31]. Ebben az esetben a sűrűség egy bizonyos ϵ sugáron belüli pontok száma. Ez alapján az algoritmus három kategóriába osztja az adatpontokat:

- Magpont: olyan adatpont, melynek egy meghatározott számúnál (μ) több pont van az ϵ sugarú környezetében.
- Határpont: olyan adatpont, amelynek μ -nél kevesebb pont van az ϵ sugarú környezetében, de ő maga egy magpont ϵ sugarú környezetébe esik.
- Zaj (outlier): olyan adatpont, amely egyike sem az előzőknek.

Működése során először felosztja az adatpontokat a három leírt típus szerint, majd a zajos pontokat figyelmen kívül hagyja. A többi pont esetén a sűrűn összefüggő részek határozzák meg a klasztereket. Előnye, hogy kiszűri a zajos pontokat és érzéketlen rájuk, illetve jól kezeli a különböző méretű osztályokat. Változó sűrűségű csoportok esetén gyengébben teljesít.

A felhasználó feladata meghatározni a μ és ϵ értékeket, illetve a megfelelő távolságfogalmat. Ehhez szintén létezik egy könyökszabály. Ennél az algoritmusnál növekvő sorrendben kell ábrázolni az egyes adatpontoknak a legközelebbi szomszédjuktól vett távolságot. Mivel sok adatpont van, a könyök simább, így az ϵ-t csak becsülni tudjuk.

Ezután a μ paramétert kell hozzáigazítani oly módon, hogy a kialakuló klaszterszám megfelelő legyen. Ez úgy történik, hogy az adott paraméterpárral lefuttatjuk az algoritmust és megnézzük a keletkezett klaszterek számát. Azt a párt választjuk, aminél megfelelő a klaszterszám.

4.3.3. Ward-féle hierarchikus klaszterező algoritmus működése

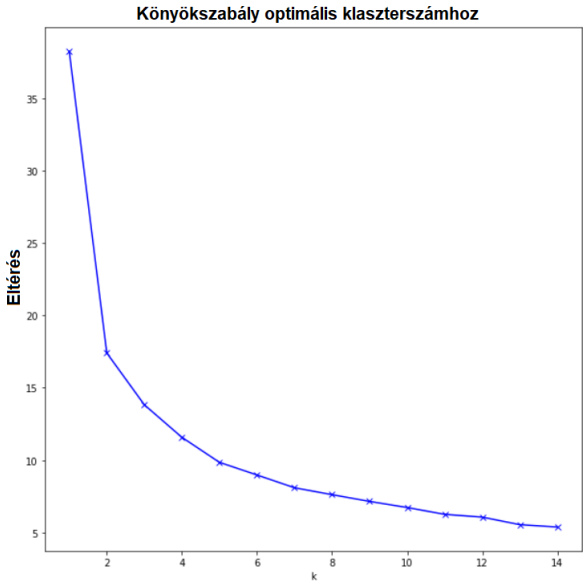
Az általunk használt hierarchikus klaszterező algoritmus agglomeratív módon hozza létre a klasztereket, azaz először minden pontot külön klaszternek tekint, majd ezeket egyesíti egy kiválasztott módszer alapján addig, amíg az összes pont egy klaszterbe nem kerül. Az iterációkat dendrogrammal lehet ábrázolni, és az alapján lehet dönteni, hány klasztert akarunk létrehozni. Gyakorlatban a dendrogramon meg kell keresni a legszélesebb vízszintes sávot, amelyben csak függőleges szakaszok vannak, és venni azok számát.

Az alapvető egyesítő módszerek a single, average, complete linkage nevet hordozzák, de mi a Ward-féle módszert alkalmaztuk. Ez a módszer az összes klaszteren belüli szórásnégyzet minimalizálására törekszik, így valamelyest hasonlít a K-közép algoritmusra.

4.3.4. K-közép algoritmus eredménye

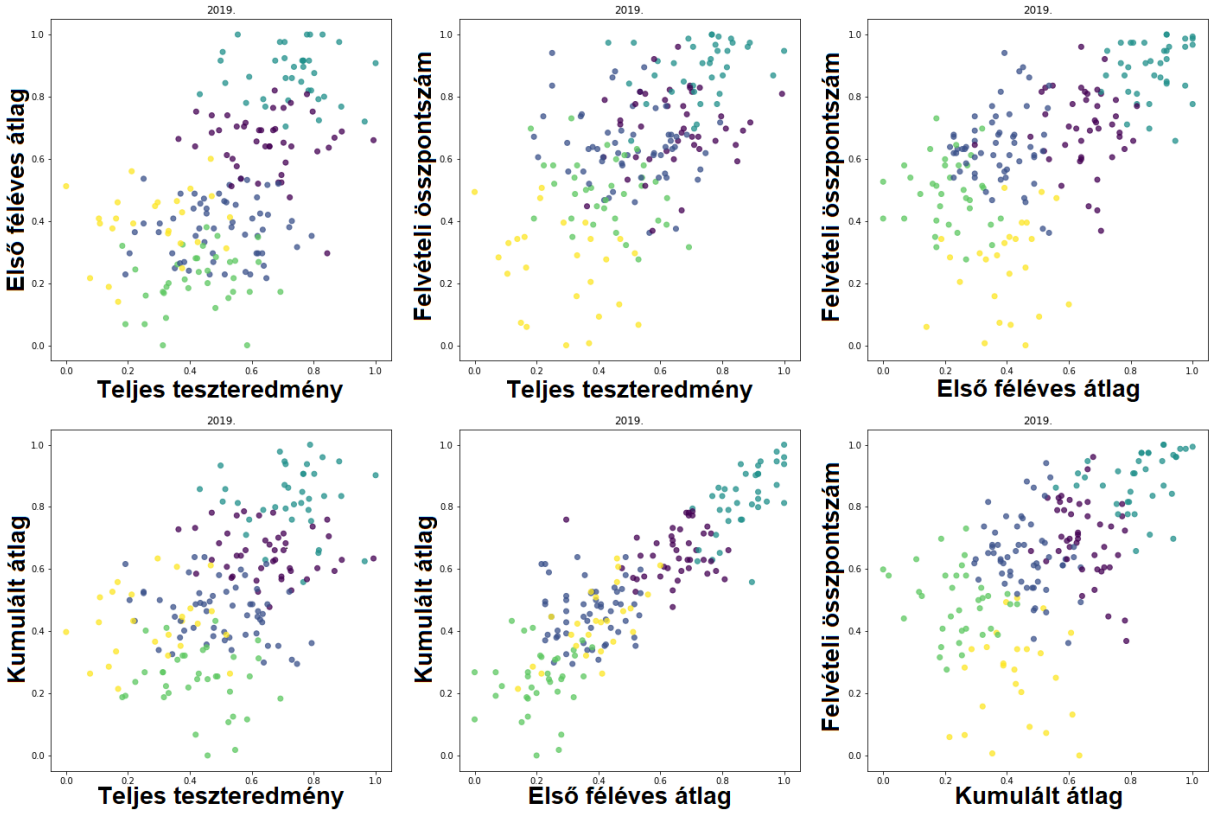
Ami a szórásdiagramokból sejthető volt, a könyökszabály alkalmazása során igazolódott be. Szinte minden próbálkozás esetén sima volt a görbe, és nem volt egyértelműen jó klaszterezés. Ezekben az esetekben a klaszterkép is elmosott volt.

A K-közép algoritmus eredményei közül kettőt emelhetünk ki. Az egyikben a 2019-es adatokból négy változó szerepelt, min-max skálázva: teljes teszteredmény, első féléves átlag, kumulált átlag és a felvételi összpontszám. A másikban ugyanez a felállás 2021-re, kivéve a kumulált átlagot, mert az megegyezett az első félévessel. Az első változatnak a könyökszabály ábrája tekinthető meg alább:



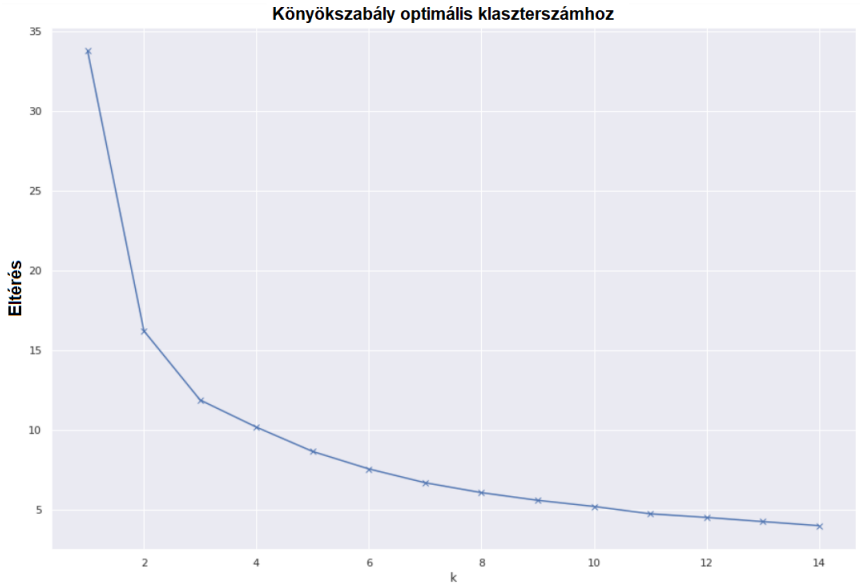
17. ábra. 2019. K-közép, 4 változós könyökszabály

Természetesen ezen az ábrán (17.) sincs teljesen egyértelmű törés, de 5 klaszter esetén mondható közel optimálisnak. Ezzel az alábbi klaszterképet generálta az algoritmus:



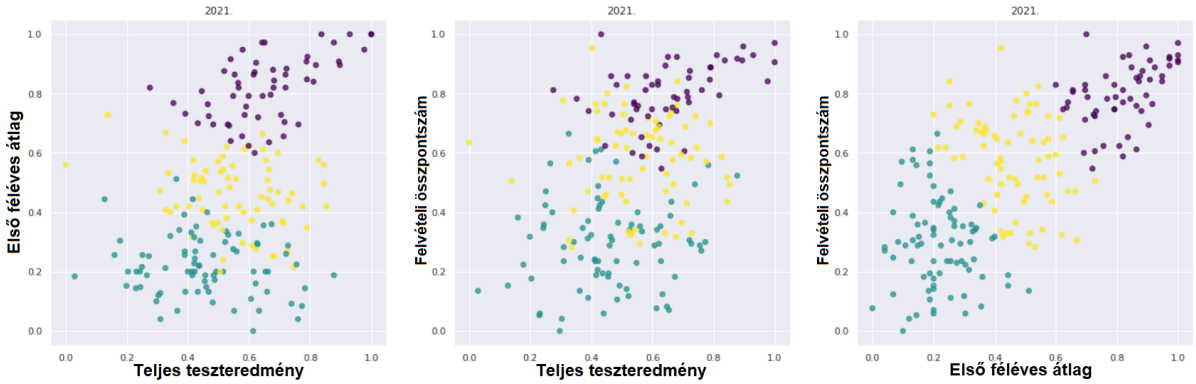
18. ábra. 2019. K-közép, 4 változós klaszterkép

A 18. ábrán lévő részábrák közül a legtisztábban a felvételi összpontszám és az első féléves átlag szórásdiagramján látszanak az osztályhatárok. Megfigyelhető az algoritmus jellegzetessége, a hasonló méretű, "gömbszerű" klaszterek kialakítása. A kékeszöld klaszter a legstabilabb, annak az adatpontjai mindig a jobb felső sarokban vannak, azaz minden változóban magas értékeik vannak. A lila hasonlóan viselkedik, csak mérsékeltebb, minden változóban átlag felettiek, de nem a legjobbak. Ezzel szemben a zöld klaszter azokból a hallgatókból áll, akik átlagosan teljesítettek a felvételi és a teszt során, de más mutatókban lecsúsztak. A kék a legbizonytalanabb, ugyanis többnyire középütt van mindig, de mélyen belemetsz más klaszterokba. A sárga klaszter pontjai minden változó esetén a skála alsó felén vannak, kivéve a kumulált átlagot, ahol középtájt összpontosulnak.



19. ábra. 2021. K-közép, 3 változós könyökszabály

A 2021-es adatokra vonatkozó 19. ábrán a könyök jobb a 17. ábrán lévő könyöknél, ugyanis 3-nál még érzékeny a törés, de utána szinte kisimul.



20. ábra. 2021. K-közép, 3 változós klaszterkép

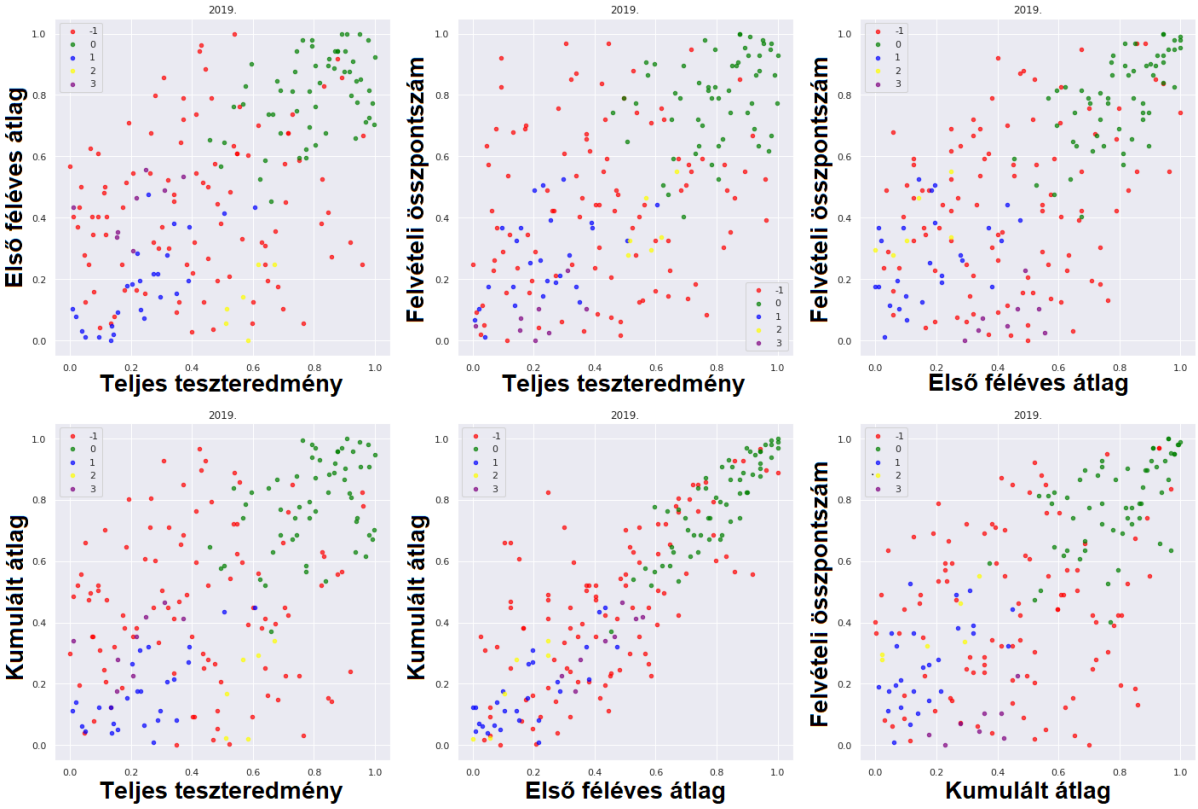
Ebben az esetben (20. ábra) még tisztább az elkülönülés az első féléves átlag és a felvételi összpontszám esetén, tehát feltehetőleg ezek jobban befolyásolták a döntést. Itt is gömbszerű klaszterek lettek, ugyanis a pontfelhő egészen homogén.

4.3.5. DBSCAN eredménye

A DBSCAN algoritmus akkor működik igazán jól, ha az adatok pontfelhője közel azonos sűrűségű részekből áll, amelyek között nagyobb távolság van. Az elkészített, és fentebb bemutatott ábrákon (4.1) nem volt jellemző az előbb említett tulajdonság. Ennek következményeként az algoritmussal nem sikerült jó klaszterezést megvalósítani.

A futtatások során az eredmények két csoportba voltak kategorizálhatóak. Az egyikben egy nagy és több elenyésző méretű klaszter jött létre, a másikban a pontok nagy része

"outlier" volt és néhány nagyobb, de továbbra is kis méretű klaszter jelent meg. Egy ilyet tekinthetünk meg alább.



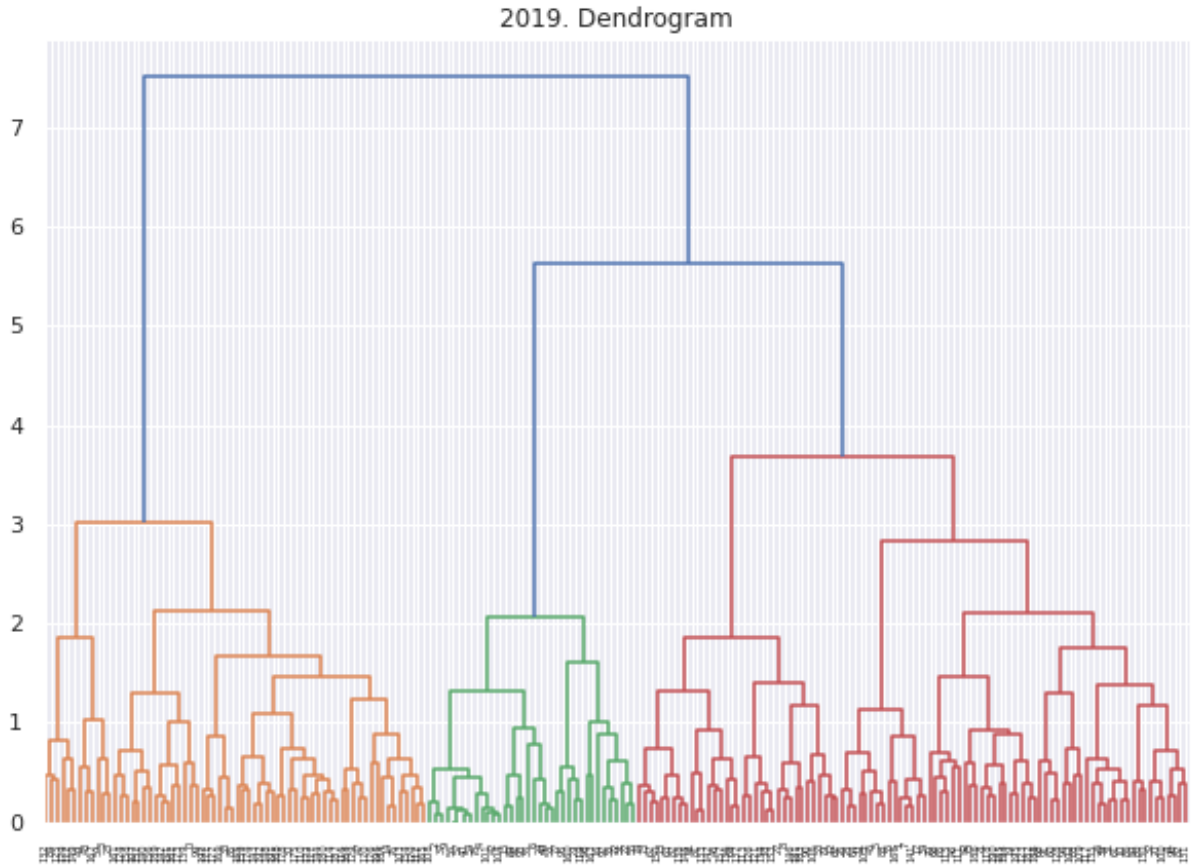
21. ábra. 2019. DBSCAN, 4 változós klaszterkép

A 21. ábrán lévő klaszterezés 2019-es adatokon készült, Mahalanobis távolságot használva és kvantilisek mentén skálázva. Ismét ugyanaz a négy változó került be: teljes teszteredmény, első féléves átlag, kumulált átlag és a felvételi összpontszám. Látszik, hogy piros pontokból van a legtöbb, amelyekhez a -1-edik klasztert rendelte az algoritmus, ezek az outlierek. A zöld és a kék klaszter még viszonylag nagy, de a többi csak pár pontot tartalmaz.

Mivel az algoritmus eredménye egyik esetben sem volt megfelelő, nem tekintjük célszerűnek a 2021-es évfolyamról készült ábrák áttekintését.

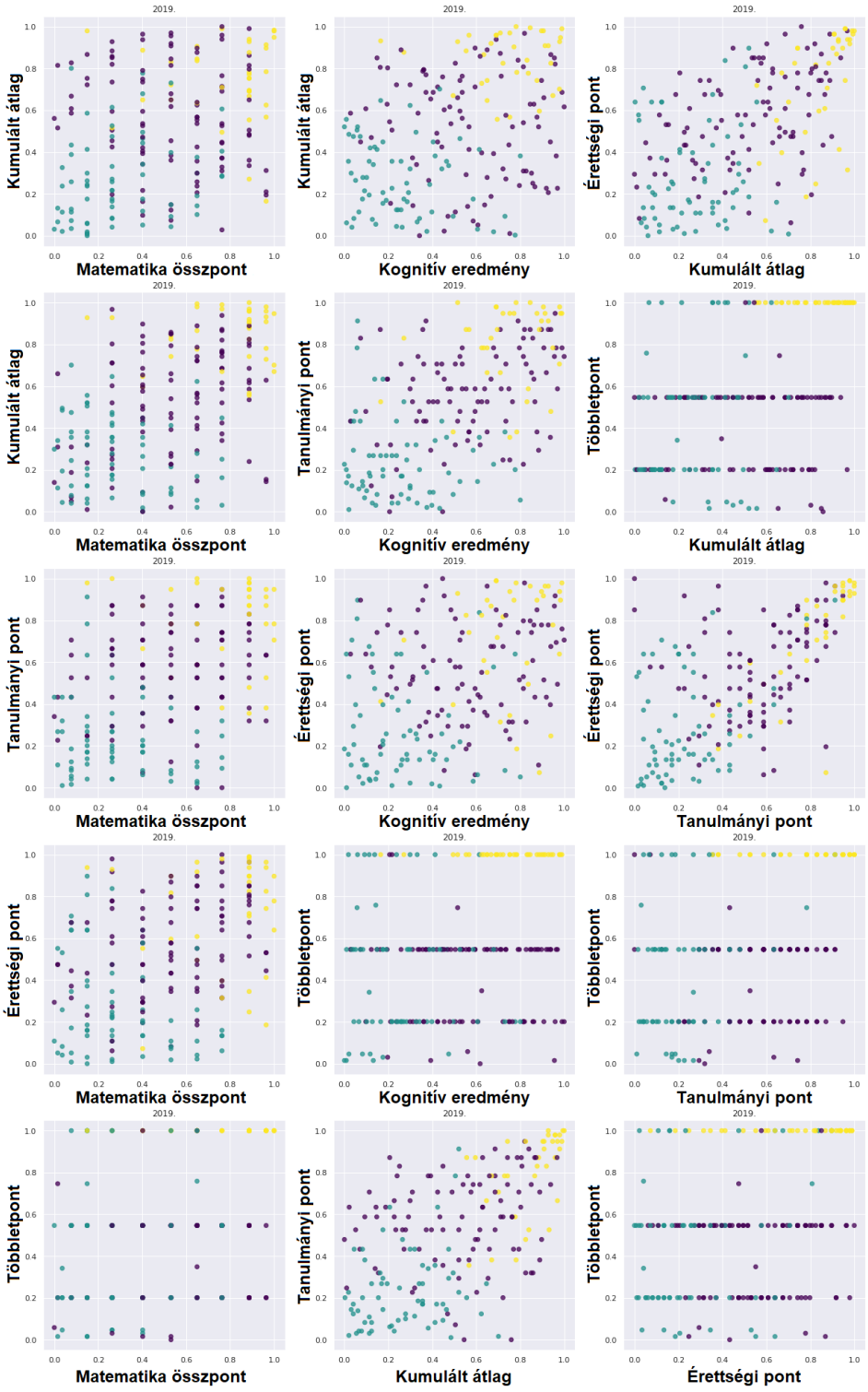
4.3.6. Ward-féle algoritmus eredménye

Mint ahogy a másik két módszernél a könyökszabályokban, itt a dendrogramokban volt nehéz eldönteni az optimális paraméter értéket, jelen esetben a klaszterszámot. Ha egy szám optimális, akkor az a síkrész, amelyben annyi függőleges vonal megy, kiemelkedően széles. Ha nehezen klaszterezhető az adat, akkor kisebbek a sávok, és azok közül a legnagyobbak közel azonos méretűek. Ezt lehet látni az alábbi, 22. ábrán:



22. ábra. Dendrogram

Itt a kettő és a három függőleges szálat tartalmazó sávok szélessége közel azonos. A választás három klaszterre esett. Kevesebb változó választásánál a klaszterek száma minden esetben legfeljebb kettő esetén lett volna optimális, így egészen sokat kellett használni: *matematika összpont, kognitív eredmény, kumulált átlag, tanulmányi pont, érettségi pont, többletpont.*



23. ábra. Ward féle algoritmus eredménye

A változók nagy száma miatt lett ilyen nagy a párosításos ábra (23.). Mindez a 2019-es adatokból van, kvantilisek mentén skálázva. Minden erőfeszítés ellenére nem alakult ki tisztább klaszterkép. Itt is a K-középhez hasonlóan a három klaszter a "jók", "közepesek" és a "gyengébbek" osztályait jelenti, de nincs vetület, amin olyan sima lenne a klaszterhatár. Ez alapján ismét nem célszerű 2021-re készült ábrát szemléltetni a dolgozatban.

5. Konklúziók a feltáró adatelemzésből

A fejezetben ismertetett szórásdiagramok adtak egy szemléletes képet az adatok egyszerű viszonyairól. Többnyire látszott, hogy a változók között pozitív korreláció volt, de nem kimondottan erős. Ez motivációt adott az adatok további, mélyebb és komplexebb vizsgálatához annak érdekében, hogy felfedezhetőek legyenek nem triviális kapcsolatok. Továbbá az ábrák összehasonlítása kimutatott egy jelentős különbséget a két év eredményei között, ami szintén indokolja a további elemzést. Szem előtt kell tartani, hogy 2021-ben a diákok már részt vettek távolléti oktatásban, és ez feltehetőleg a változások kiváltó oka lehetett.

A folyamatábrák vizsgálata fényt derített több érdekességre. Például a 8. ábrából azt a következtetést vonhattuk le, hogy míg a bemeneti eredmények viszonylag erősen befolyásolták az első néhány megmérettetés eredményét, a további eredmények már nagyobb részben nem függtek tőlük. Ez az $A1$ és $A2$ tárgyak közötti átmenetek alapján látszik tisztán. Ez kívánt eredmény, ugyanis az egyetem szerepe nem lenne jelentős, ha csak a felvételi eredményeken múlna a további sikeresség.

A Sankey-diagramok szintén fényt derítettek a két év eredményességeinek különbségére. Ennek a leglátványosabb példája a 15. és 16. ábra, ahol a tanulmányi pontok eloszlásai között óriási különbség van. A két évben az $A1$ jegyek eloszlási között is érdemi a különbség, de az elvárásokkal szemben 2021-ben romlottak. Ennek több oka lehet, többek közt a tárgykövetelmény szigorítása.

Összességben a folyamatábrákból az a következmény vonható le, hogy a bemeneti eredményeknek van hatása a továbbhaladásra, de idővel gyengül. Továbbá a változók értékkészletének két szélén lévő hallgatók eredményei könnyebben követhetők, míg a közepes hallgatók esetében nagyobbak a fluktuációk.

A szórásdiagramok eredményeinek elemzése alapján a klaszterezéstől nagy áttörés nehezen volt elvárható. A további kísérletek során bebizonyosodott, hogy a birtokunkban lévő adatok nehezen klaszterezhetőek. Ennek ellenére a K-közép algoritmussal sikerült elérni egészen tiszta klaszterhatárokat, és azt figyeltük meg, hogy ezek az első féléves átlag és a felvételi összpontszám esetében voltak a legtisztábbak. A DBSCAN algoritmushoz egyáltalán nem volt megfelelő az adat, a Ward-féle módszer szintén gyengébben teljesített a K-középnél.

6. Prediktív analitika

6.1. Modellek és metodológia

Következő lépésként azt vizsgáltuk, hogy a két évben külön a bemeneti adatok alapján mennyire pontosan lehet előrejelezni egyes teljesítménymutatók értékét, illetve, hogy a predikciókra nézve a különböző bemeneti változók, más néven attribútumok, milyen és mekkora szereppel bírnak. Ehhez többféle modellen többféle *gépi tanuláson alapuló osztályozó és regressziós algoritmus* használatára volt szükség. A kutatás során kétféle eredmény alaposabb vizsgálatára összpontosítottunk mindkét évben: a "Matematika A1a - Analízis" tárgyból kapott érdemjegyre illetve az első félév végén megállapított kumulált tanulmányi átlagra. Ezen feladatok a felügyelt gépi tanulási problémák közé esnek, ahol a kitüntetett célváltozót szeretnénk a többi bemeneti változóval előrejelezni [8]. Ekkor a gépi tanulás során a teljes adathalmazt tanító és teszthalmazra bontjuk, a tanító adathalmazon betanítjuk az algoritmusokat úgy, hogy a tanítóhalmazbeli adatok célváltozóját minél pontosabban prediktálják, majd valamilyen metrika szerint kiválasztjuk a betanított algoritmusok közül a legjobbat. Az optimális osztályozó kiválasztásához többnyire *keresztvalidációt* alkalmazunk. Keresztvalidáció során a tanítóhalmazt felbontjuk K egyenlő részre, melyek közül az egyiket kinevezzük validációs halmaznak. Ezt követően az algoritmusokat betanítjuk a maradék K-1 részen, amelyeket aztán a validációs halmazon kiértékelünk. Ezt összesen K alkalommal ismételjük, mindig másik részt választva validációs halmaznak, majd azt az algoritmust dedikáljuk a legjobbnak, amelynek az aggregált teljesítménye a K darab iteráció során a legjobb. Az így kapott osztályozót még visszamérjük a teszthalmazon is az általánosítóképesség ellenőrzése végett.

A két vizsgált probléma közül az előbbi alapvetően egy osztályozási probléma öt osztállyal, míg az utóbbi egy regressziós feladat. Az előbbi feladathoz ötféle algoritmust, *Gradient Tree Boosting-ot, Naive Bayes-t, logisztikus regressziót, SVM-et* és *lineáris regressziót* használtunk, az utóbbihoz pedig lineáris regressziót és Gradient Tree Boosting-ot (ezen algoritmusok működése és optimalizálása a következő fejezetben lesz ismertetve). Azért esett a választás ezen algoritmusokra, ugyanis számos egyetemi teljesítményt és lemorzsolódást vizsgáló tanulmány során teljesítettek kiváló eredménnyel [2–5]. Mivel azonban viszonylag kevés adat állt a rendelkezésünkre, ezért az érdemjegyek prediktálásánál az adatrekordok esetén a pontos érdemjegy helyett érdemjegycsoportok előrejelzésére koncentráltunk. A matematika érdemjegycsoportok prediktálására így kétféle modell került felvázolásra: egy *3 csoport modell*, illetve egy *2 csoport modell*.

A 2 csoport modell esetén a két csoportot a {5,4,3} illetve {2,1} osztályok adják, míg a 3 csoport modellnél az osztályok {5,4}, {3,2} illetve {1} módon alakultak. Az előbbi esetben az osztályok intuitívan a lemorzsolódási veszélyeztetettség szerint formálódtak, míg az utóbbiban egy általánosabb "jól teljesítő", "rosszul teljesítő", "lemorzsolódott" csoporthármast kívántunk elérni. Mindkét modell esetén külön vizsgáltuk a teljesítményt összes hallgatóra vonatkozólag illetve szétbontva vegyészmérnök és biomérnök hallgatókra egy-

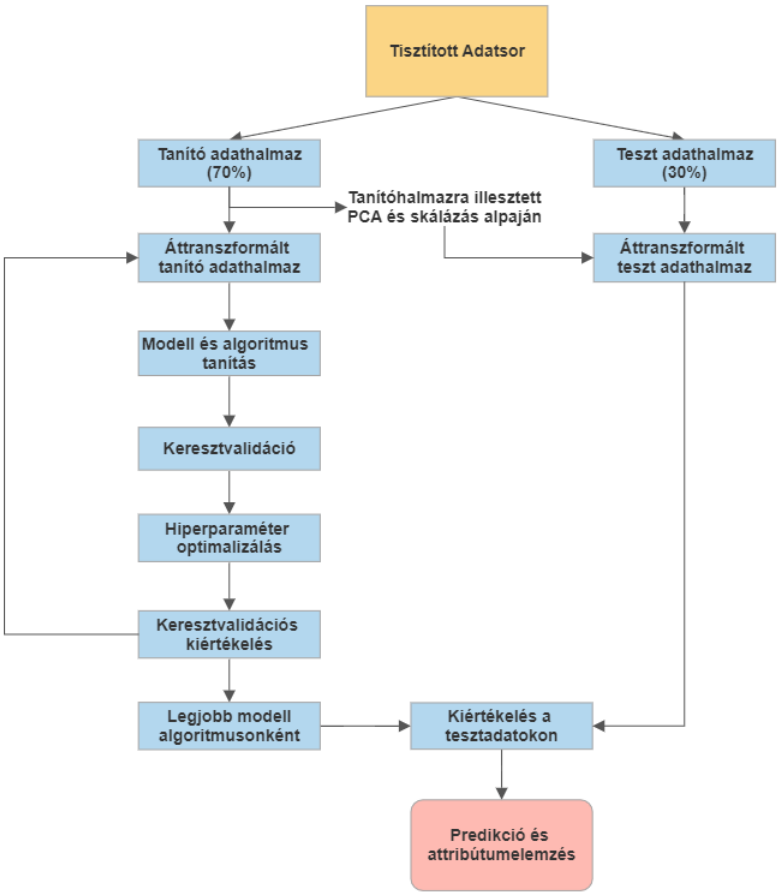
aránt (a környezetmérnökökről nem állt rendelkezésre elég adat, így őket a szakonkénti bontásban kihagytuk).

A használandó osztályozó algoritmusok közül a Naive Bayes és Gradient Tree Boosting algoritmusok képesek kezelni a többosztályos feladatokat, a lineáris regresszió azonban folytonos célváltozóérték prediktáláshoz használható, így ott a prediktált értéket kerekítettük a legközelebbi címkeértékhez. A logisztikus regresszió és az SVM alapjáraton csak bináris osztályozásra alkalmas, így a náluk *One-vs-Rest* elvű osztályozást használtunk. Az elv lényege, hogy minden osztály esetén képezünk egy új virtuális osztályt, amely az összes többi osztályt tartalmazza, majd minden ilyen osztálypár esetén a bináris osztályozó meghatározza a kérdéses adatrekord esetén az osztályba tartozási valószínűségeket. Végül azt a címkét prediktáljuk a rekordnak, amely osztály esetén a legnagyobb ez a valószínűség.

Az egyes modellek és almodellek esetén a tanítás előtt főkomponens analízist ("Principal Component Analysis", röviden PCA) is alkalmaztunk. Az eljárás lényege, hogy az adatpontokat egy kisebb dimenziós térre vetítjük le oly módon, hogy a változók közötti variancia minél nagyobb részét tartsák meg, így minimalizálva az információveszteséget [10]. Az új változók, amelyeket főkomponenseknek nevezünk, a kisebb dimenziós térben az eredeti változók tapasztalati kovarianciamátrixának sajátvektorai lesznek. PCA használata során az algoritmusok gyorsabban tanulnak a kisebb dimenziószám miatt, és olykor jobb eredményt is érnek el. Jelen kutatásban másrészt azért is döntöttünk a PCA alkalmazása mellett, mert a korábbi, ugyanezen problémakört vizsgáló, általunk átnézett tanulmányok nem használtak PCA-t még nagyobb attribútum szám esetén sem, ugyanakkor mi a korai tanító-tesztelési fázisnál azt tapasztaltuk, hogy az áttranszformált adatokon jobban teljesítenek az osztályozó algoritmusok, így potenciálisan megéri alkalmazni.

Implementálás

A vizsgálatokat Python-ban végeztük el az *Sklearn* [6] csomag használatával. A *jegycsoportok* prediktálására épített modellezési struktúrát a 24. ábra mutatja. Az egyes csoportmodelleknél a teljes adathalmazt felbontottuk tanító- és teszthalmazra 70-30 arányban, majd a tanítóhalmazra illesztett PCA modellt alkalmaztuk a teszthalmazra is, ahol a főkomponensek számát minden modellnél 2 és 8 között iterálva változtattuk 2-es lépésközzel. Ezt követően a változókat kvantilis alapú [0,1] uniform skálázásnak vetettük alá a tanítóhalmazra illesztett tapasztalati eloszlásfüggvények szerint, valamint az érdemjegycsoportok prediktálásához a 2 és 3 csoport modelleknél a célváltozó-értékeket rendre 1, 0-ra illetve 3, 2, 1-re módosítottuk.



24. ábra. A jegycsoportokra irányuló modellezési struktúra sematikus ábrája

Valamennyi algoritmus hiperparamétereinek optimális megválasztására 5-szörös keresztvalidációt alkalmaztunk, ahol törekedtünk arra, hogy az adatrekordok címkéjének eloszlása egyenletes legyen a felosztott részek között. A keresztvalidációnál használt, jóságot mérő metrikának a kiegyensúlyozott pontosságot ('*balanced accuracy*') választottuk, amelynek képlete az alábbi:

$$\text{balanced-accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{2}$$

Ezen metrika alapvetően bináris, pozitív-negatív osztályú osztályozási problémákhoz alkalmas, de többosztályos osztályozás esetén is használható, ahol az egyes osztályokhoz tartozó *TP, FP, FN, TN* értékek által kiszámított kiegyensúlyozott pontosságok számtani közepét nézzük. A képlet egyes jelölései két valamint több osztály esetén:

- **TP** a pozitívnak (osztálybelinek) osztályozott, valóban pozitív (osztálybeli) adatrekordok száma
- **FP** a pozitívnak (osztálybelinek) osztályozott de valójában negatív (nem osztálybeli) adatrekordok száma
- **FN** a negatívnak (nem osztálybelinek) osztályozott de valójában pozitív (osztálybeli)

adatrekordok száma
- **TN** a negatív (nem osztálybelinek) osztályozott, valóban negatív (nem osztálybeli) adatrekordok száma

A választott metrika mellett meghatároztuk a legjobb algoritmusokat, amelyeket a teszthalmazon visszamértünk, valamint kiértékeltük az ily módon választott modellek esetén az egyes változók fontosságát is.

Az *első féléves kumulált átlag* predikciójánál a tisztított adathalmazt ugyanúgy 70-30 arányban osztottuk fel. Adattranszformálásra csak kvantilis alapú skálázást alkalmaztunk, PCA-t nem, majd valamennyi regresszor hiperparamétereinek optimalizálása hasonlóan 5-szörös keresztvalidációval történt. Az illesztett legjobb modelleknél mindkét évben feljegyeztük a reziduális tagok[1] szórásdiagramját, az egyes változók prediktív erejét, valamint az alábbi statisztikák értékét [9]:

- R^2: A modell hatásfokát mérő mutató, értéke 0 és 1 közötti (nagyon gyengén teljesítő modell esetén negatív értéket is felvehet). Azt mutatja meg, hogy az adatok között fellépő variancia mekkora részét képes a regressziós modell megmagyarázni:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \tag{3}$$

  ahol y_i és $\hat{y}_i$ rendre az i-edik adatrekord valódi és prediktált célváltozó-értéke, $\overline{y}$ pedig a valódi célváltozó-értékek átlaga.

- MAE: Átlagos abszolút eltérés (A keresztvalidáció során ez volt a használt mérőszám is):

$$MAE = \frac{1}{n} \sum_i^n |y_i - \hat{y}_i| \tag{4}$$

- $RMSE$: Gyököt vont átlagos négyzetes eltérés:

$$RMSE = \sqrt{\frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2} \tag{5}$$

6.2. Osztályozó algoritmusok és optimalizálásuk

6.2.1. Lineáris regresszió

A lineáris regresszió jellegéből adódóan alapvetően folytonos célváltozó prediktálására alkalmas, ugyanakkor kategorikus, ordinális címkéjű adatok osztályozására is fel lehet használni. Az algoritmus lényege, hogy a magyarázóváltozók mindegyikéhez egy-egy súlyt rendelünk, majd az adatpont attribútumértékeinek vesszük a súlyokkal való súlyozott

[1] A reziduális tagok a célváltozó-értékeknek a prediktált értékektől vett különbségei.

összegét, és esetleg egy bias tagot hozzáadva az így kapott szumma lesz a prediktált címkeérték. A cél a tanítás során a súlyok optimális megtalálása, miközben a tanítóhalmazon minimalizáljuk a célváltozóérték és a prediktált érték közötti négyzetes hibát.

Mivel az algoritmus az implementálása végett a legkisebb négyzetek elvét alkalmazza, amellyel az optimális megoldás analitikusan elérhető, ezért nem történt hiperparaméter-optimalizáció.

6.2.2. Naive Bayes

A Naive Bayes algoritmus működése mögött álló alapelv az, hogy feltesszük az attribútumok feltételes függetlenségét, amennyiben a célváltozó értéke ismert. Osztályozás során azt vizsgáljuk, hogy mely címkeérték mellett a legnagyobb a valószínűsége annak, hogy az adott adatrekord attribútumai éppen a felvett értékeket kapták. A megfelelő valószínűségeket a tanítóhalmazbeli adatok attribútumértékeinek különböző címkék melletti relatív gyakoriságai adják.

Az algoritmus a jellegéből adódóan nem igazán optimalizálható, így a Naive Bayes esetén nem történt keresztvalidáció.

6.2.3. Gradient Tree Boosting

A Gradient Boosting algoritmus egy *ensemble* típusú osztályozó, amelynek lényege, hogy sok gyenge teljesítményű prediktor (*"weak learner"*) eredményét felhasználva hoz létre egy erős prediktort [1]. Gradient Tree Boosting esetén a gyenge prediktorok *döntési fák*, amelyek lefelé irányított, legtöbbször bináris fák, továbbá minden belső csúcsban egy attribútumra vonatkozó feltétel szerepel, a levelek pedig valamilyen célváltozó-értékkel címkézettek. Az adatrekordok osztályozása intuitívan a fán való végigvezetéssel történik, végül azt a címkét prediktálva nekik, amilyen címkéjű levélbe jutottak.

A boosting eljárás során minden fázisban egy új döntési fát építünk, amely megpróbálja az előző fázisban épített fa hibáit csökkenteni az úgynevezett reziduálisokra építve. A cél egy erős prediktorként funkcionáló döntési fa létrehozása ráerősítések sorozata révén. Az algoritmus alkalmas osztályozási és regressziós problémák megoldására is, előbbi esetben a cél a levelekben az adatrekordok címke szerinti homogenitásának maximalizálása, utóbbinál pedig a szórás minimalizálása az egy levélbe kerülő rekordok célváltozójára nézve.

A keresztvalidálás során optimalizált paraméterek:
- Tanulóráta (0.01 és 5 között változtatva 0.1-es lépésközzel)
- Boosting fázisok száma (5 és 100 között változtatva 5-ös lépésközzel)
- Facsúcsokban használt vágási feltétel (négyzetes hiba, Friedman MSE)
- Fák maximális mélysége (3,4,5 és 6 között változtatva)

6.2.4. Logisztikus regresszió

A logisztikus regresszió alapvetően bináris osztályozási problémák megoldására alkalmas, de kiterjeszthető többosztályos feladatok megoldására is. Lényege, hogy a lineáris regresszióhoz hasonlóan az adatrekord attribútumértékeinek súlyozott összegét használjuk egy szigmoid[2] függvény bemeneteként, amelyet a függvény leképez a $(0, 1)$ intervallumra, és amennyiben az output 0.5-nél nagyobb, úgy a pozitív osztályba soroljuk az adott rekordot, különben a negatívba. A cél nyilván itt is az optimális súlyok megtalálása, amellyel a félreosztályozási hibát minimalizáljuk.

A keresztvalidálás során optimalizált paraméterek:
- Regularizációs paraméter (0.05 és 5 között változtatva 0.05-ös lépésközzel)
- Önoptimalizálási módszer ("SAG", "SAGA")

6.2.5. SVM

Az SVM (Support Vector Machine) algoritmus a logisztikus regresszióhoz hasonlóan egy lineárisan szeparáló hipersíkot akar meghatározni [7]. Lényege, hogy különböző magfüggvények segítségével az adatrekordokat egy magasabb dimenziós térbe képzi le, ahol olyan szeparáló hipersíkot keres, amely maximalizálja a vele párhuzamos hipersíkok által meghatározott olyan térrészt, amely adatpontot nem tartalmaz. A cél a megfelelő magfüggvény és az optimális hipersík megtalálása, amellyel a különböző címkéjű adatpontok lineárisan szeparálhatóak.

A keresztvalidálás során optimalizált paraméterek:
- Regularizációs paraméter (0.1 és 5 között változtatva 0.1-es lépésközzel)
- Magfüggvény (lineáris; legfeljebb 3-ad fokú polinomiális; radiális bázisfüggvény)

[2]A szigmoid függvény: $\sigma(z) = \frac{1}{1+e^{-z}}$, ahol $z = x_1w_1 + x_2w_2 + \cdots + x_nw_n + b$ a súlyozott összeg.

7. Modellek kiértékelése

			Osztályozó algoritmusok				
			Grad. Boost.	Naive B.	Log. reg.	SVM	Lin. reg.
3 csoport	Összesített	2 PC	66.67	62.67	52.00	58.67	61.33
		4 PC	**70.67**	60.00	57.33	53.33	65.33
		6 PC	65.33	62.67	54.67	68.00	64.00
		8 PC	64.00	68.00	53.33	62.67	64.00
	Szakonként	2 PC	78.71	80.36	73.85	67.24	80.36
		4 PC	75.41	80.36	47.80	65.59	78.71
		6 PC	**82.02**	80.36	75.47	67.21	82.02
		8 PC	80.36	78.71	37.42	75.44	78.71
2 csoport	Összesített	2 PC	59.42	69.57	69.57	72.46	69.57
		4 PC	59.42	71.01	**73.91**	69.57	73.91
		6 PC	63.77	69.57	73.91	71.01	73.91
		8 PC	68.12	69.57	71.01	73.91	72.46
	Szakonként	2 PC	67.31	67.31	65.69	62.41	64.03
		4 PC	64.00	64.00	67.31	67.31	63.97
		6 PC	64.03	65.69	64.07	65.65	67.27
		8 PC	64.07	62.51	57.49	67.34	**68.96**

2. táblázat. A 2019-es adatsor eredményei

Először a **jegycsoportok** prediktálásának eredményeit ábrázoljuk. A 2. táblázatban a 2019-es adatokra optimalizált modellek teljesítménye látható, ahol az első három oszlop a modellt és a használt főkomponensek (PC) számát mutatja, míg a jobb oldali öt oszlop az optimalizált algoritmusok teljesítményét szemlélteti. A szakonkénti bontásban szereplő értékek a biomérnöki és vegyészmérnöki adatokon kapott értékek átlagai súlyozva az egyes szakokon tanuló hallgatók számával. Az algoritmusok többségének hatásfoka 60-80% közé tehető. A 3 csoport modellek esetén a Gradient Tree Boosting algoritmus, míg a 2 csoport modellek esetén a regressziós algoritmusok teljesítettek a legjobban, ugyanakkor a 2 csoport modellek teljesítménye többségében alulmúlja a 3 csoport modellekét. Mivel a 2 csoport modellben az osztályok eloszlása kiegyensúlyozottabb, ez arra enged minket következtetni, hogy a 2-es és 3-as érdemjegyet szerzők között a bemeneti adatok tekintetében nincsenek nagy különbségek.

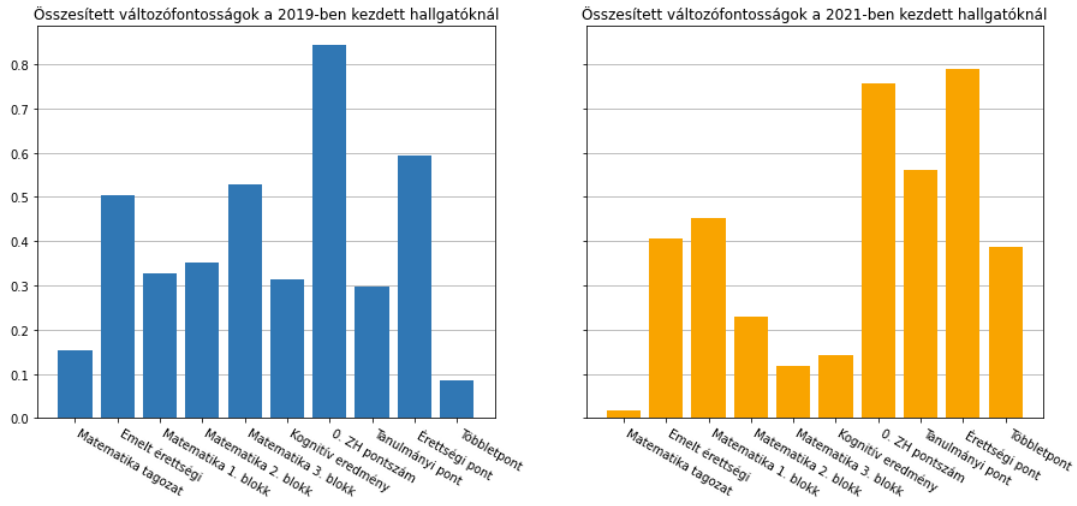
A 3. táblázat a 2021-es adatokon optimalizált algoritmusok eredményét szemlélteti az előző ábráéval megegyező metodológia szerint. Ezen adathalmazon az osztályozók teljesítménye jobbnak mondható, mint a 2019-es adatsoron, átlagosan a legeredményesebb algoritmusnak a Naive Bayes nevezhető, amely a szakonkénti bontású 3 csoport modellen ért el minden főkomponensszám mellett 80% feletti teljesítményt, ugyanakkor a 2 csoport

modell esetén a regressziós és SVM algoritmusok is 80% közeli vagy afölötti eredményt értek el. A 2019-es eredményekkel ellentétben a 2021-es adatokon a 2 csoport modellek értékei jobbak, mint a 3 csoport modelleké, ugyanakkor nem szignifikánsan.

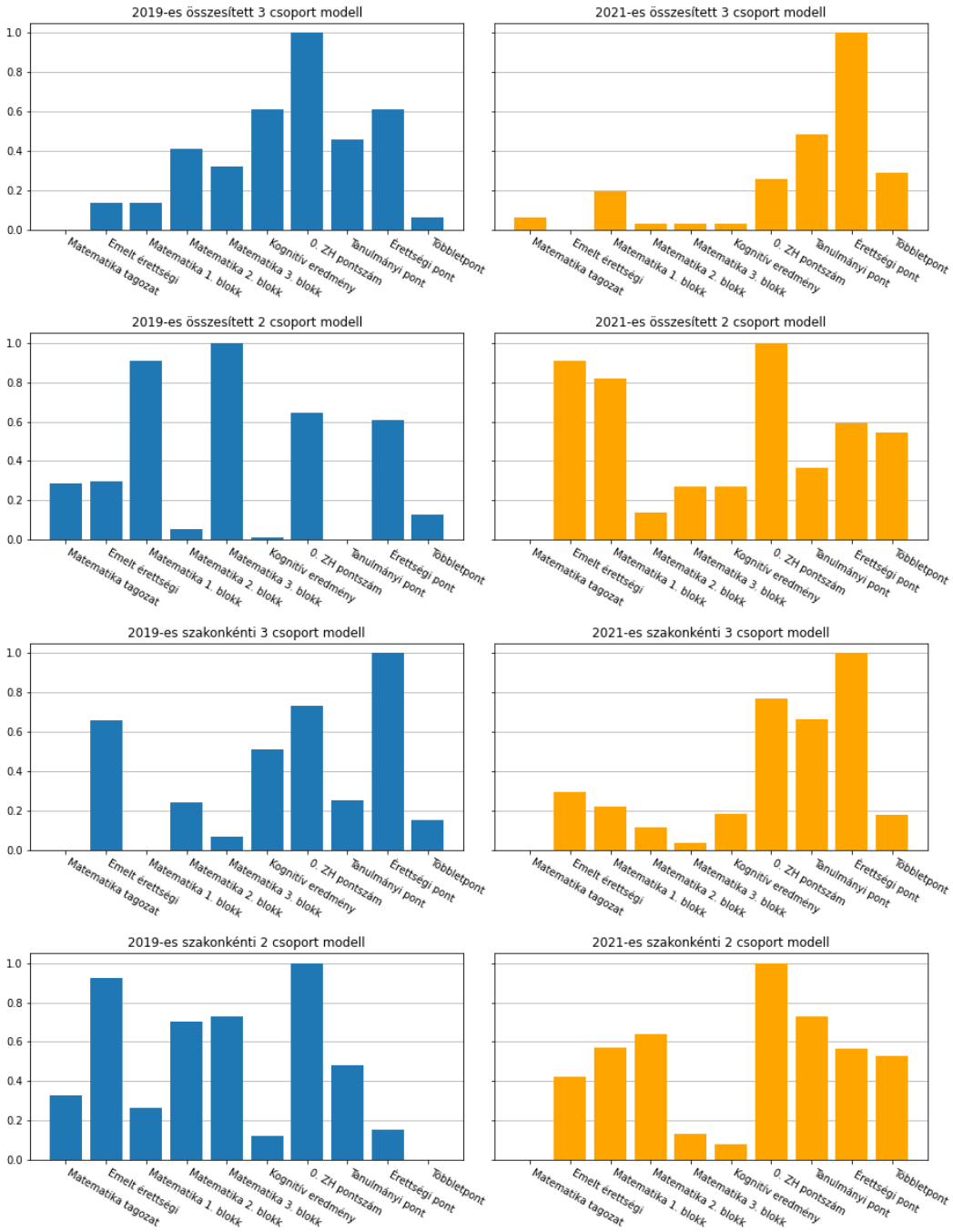
			Osztályozó algoritmusok				
			Grad. Boos.	Naive B.	Log. reg.	SVM	Lin. reg.
3 csoport	Összesített	2 PC	54.24	**80.39**	76.29	78.81	60.92
		4 PC	54.89	72.05	72.13	75.79	70.40
		6 PC	66.24	76.44	70.04	76.94	70.62
		8 PC	63.00	75.29	68.18	73.06	62.07
	Szakonként	2 PC	64.84	81.89	59.23	64.77	49.35
		4 PC	55.55	**83.07**	66.69	66.66	67.91
		6 PC	66.69	82.85	66.65	53.67	53.31
		8 PC	59.26	82.96	68.51	62.95	55.53
2 csoport	Összesített	2 PC	69.45	67.89	69.77	69.77	67.89
		4 PC	72.74	72.90	78.23	76.51	77.91
		6 PC	67.89	72.90	79.80	79.80	76.19
		8 PC	68.05	74.62	79.96	**83.24**	79.63
	Szakonként	2 PC	60.23	77.41	72.87	73.01	79.68
		4 PC	63.92	74.86	72.59	69.60	78.27
		6 PC	67.61	77.84	81.08	76.42	**81.53**
		8 PC	68.32	77.84	77.41	69.75	79.68

3. táblázat. A 2021-es adatsor eredményei

A két év modelljeinek összesített változófontosságait a 25. ábra, míg az egyes modellekhez tartozó attribútumszignifikanciákat a 26. ábra szemlélteti.



25. ábra. A változók összesített fontossága a két évben

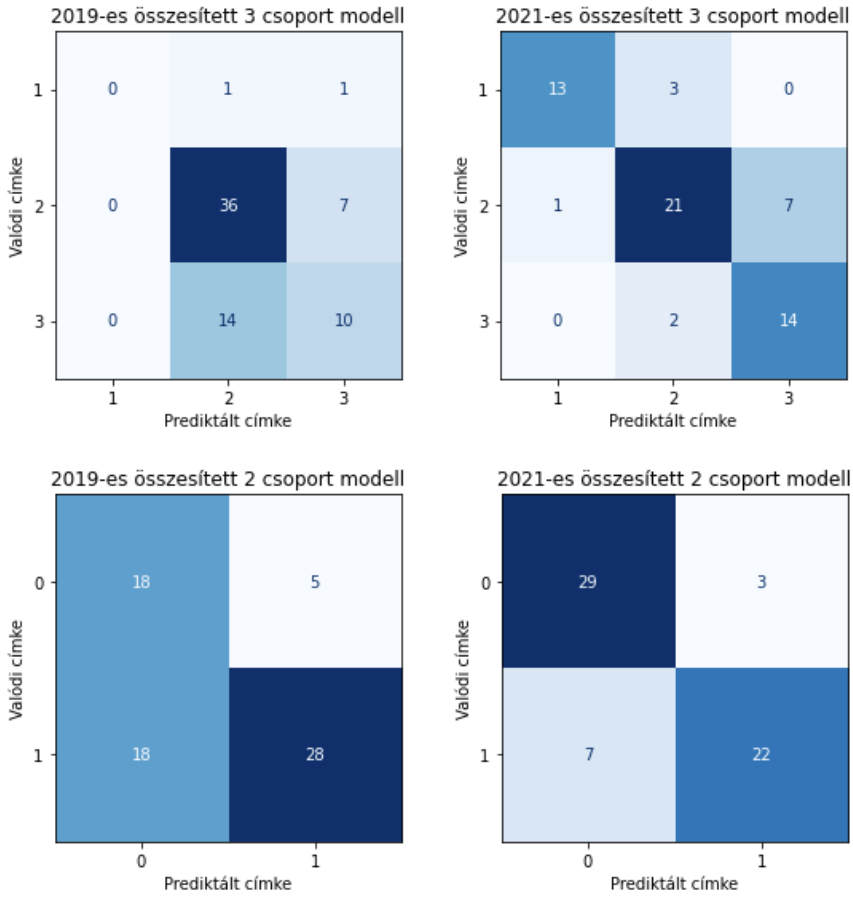


26. ábra. Az attribútumok prediktív ereje a legjobban teljesítő algoritmusoknál

A diagramokon szereplő értékek az egyes algoritmusoknál megállapított változófontosságok min-max skálázott értékei, amelyeket regressziós algoritmusoknál az attribútumokhoz rendelt súlyokból, a többi osztályozónál pedig az Sklearn *'inspection'* csomagjának segítségével nyertünk ki. Az összesített ábrán az egyes évek legfontosabb attribútumai a megfelelő évek különböző modelljein is többnyire jelentős szignifikanciával bírnak, a többi változó fontossága viszont modellenként eltérő. A 0. ZH pontszám illetve az érettségi pont prediktív ereje mindkét évben kiemelkedő. A 2019-es évben az emelt matematika érettségi megléte és a matematika-nyelvi teszten elért eredmény is nem elhanyagolható

szignifikanciával bírt, viszont 2021-ben ezen tényezők prediktáló ereje csökkent, ugyanakkor az elért tanulmányi- és többletpontok jelentősége a matematika jegyre nézve közel kétszeresére nőtt.

A félreosztályozások jellegéről több információval bírnak a *tévesztési mátrixok*, amelyek a prediktált és valódi címke vonatkozásában ábrázolják az egyes adatrekordok számát. A 27. ábrán az összesített modelleknél legjobban teljesítő algoritmusok tévesztési mátrixai láthatóak, ahol a 3 csoport modellnél a 3, 2, 1 címkék rendre az {5,4}, {3,2}, {1} osztályokat, míg a 2 csoport modellnél az 1 és 0 címkék pedig az {5,4,3} és {2,1} osztályokat reprezentálják. Egy optimális osztályozó esetén csak az átlóban vannak nem nulla elemek, hisz akkor egyezik meg a prediktált és a valódi címkéje az adatoknak. Az ábrán jól látható módon míg 2021-ben a félreosztályozás mértékében az egyes osztályok között nincsenek számottevő különbségek, addig 2019-ben a 3 csoport modellnél a {5,4}-ös osztálybeli adatrekordok nagyobb része {3,2}-as osztályba lett besorolva (14 darab hallgató a 24-ből), valamint a 2 csoport modell esetén magas a hamisan 0-nak, azaz lemorzsolódásban veszélyeztetettnek osztályzott rekordok aránya. Ez azt mutatja, hogy az illesztett modellek sok esetben alulbecsülik a 2019-es hallgatók teljesítményét, amelynek az egyik oka az lehet, hogy az osztályok között vékonyak a határok a bemeneti adatok tekintetében.



27. ábra. Tévesztési mátrixok az összesített modellek esetén

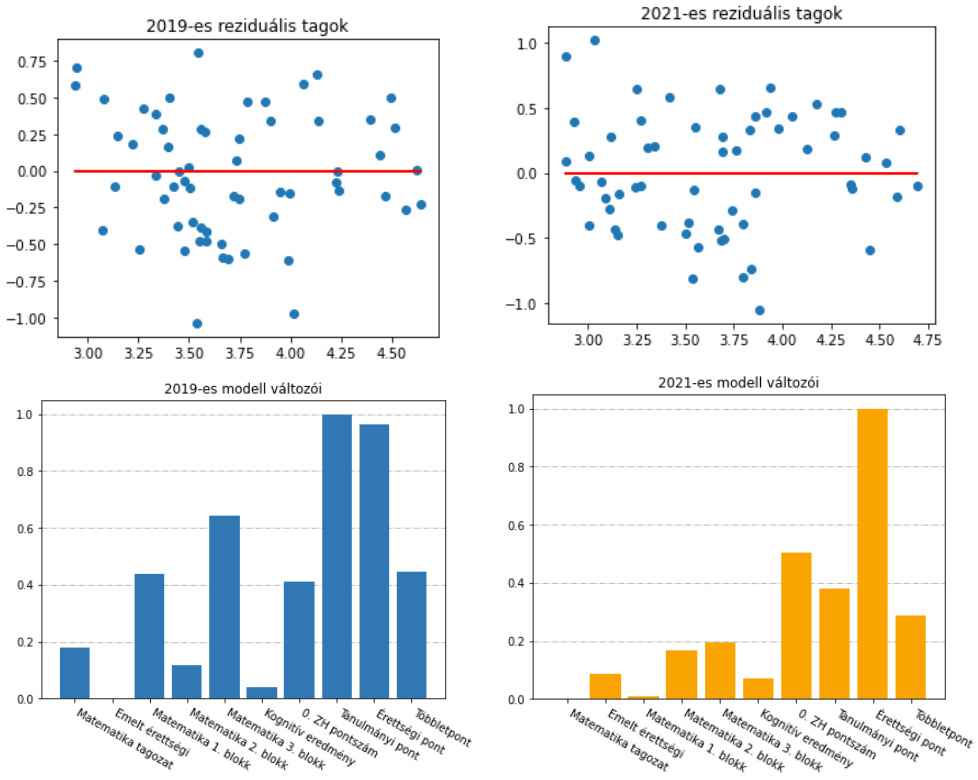
Az **első féléves kumulált átlag** lineáris regresszióval való prediktálásának eredményei (statisztikák, reziduális szórásdiagramok és az egyes változók prediktív ereje egymáshoz viszonyítva) a 28. ábrán láthatóak.

2019

- R^2: 0.5462
- MAE: 0.3455
- $RMSE$: 0.4165

2021

- R^2: 0.5646
- MAE: 0.3710
- $RMSE$: 0.4415



28. ábra. Lineáris regresszió eredménymutatói a két évben

Statisztikák tekintetében mindkét évben hasonló eredményeket kaptunk, amely azt mutatja, hogy a két évben hasonló hatékonysággal lehet a kumulált átlag eloszlását modellezni lineáris modellekkel, ugyanakkor csak közepes hatékonysággal. Ami kiemelendő, az a reziduális tagok eloszlása, amely mindkét évben hasonló alakzatot vesz fel, bár 2019-ben kicsivel nagyobb szórással. A változók prediktív erejét tekintve itt is fennáll az az érdemjegycsoportok prediktálásánál megállapított jelenség, miszerint a matematika-nyelvi teszten elért eredmény 2019-ben magas szignifikanciával bírt, azonban 2021-ben ez a szignifikancia drasztikusan csökkent. Az érettségi pont mindkét évben igen fontos determináló tényező, ugyanakkor 2019-ben a hozott tanulmányi pontok is nagy szerepet játszanak az első féléves kumulált átlag meghatározásában.

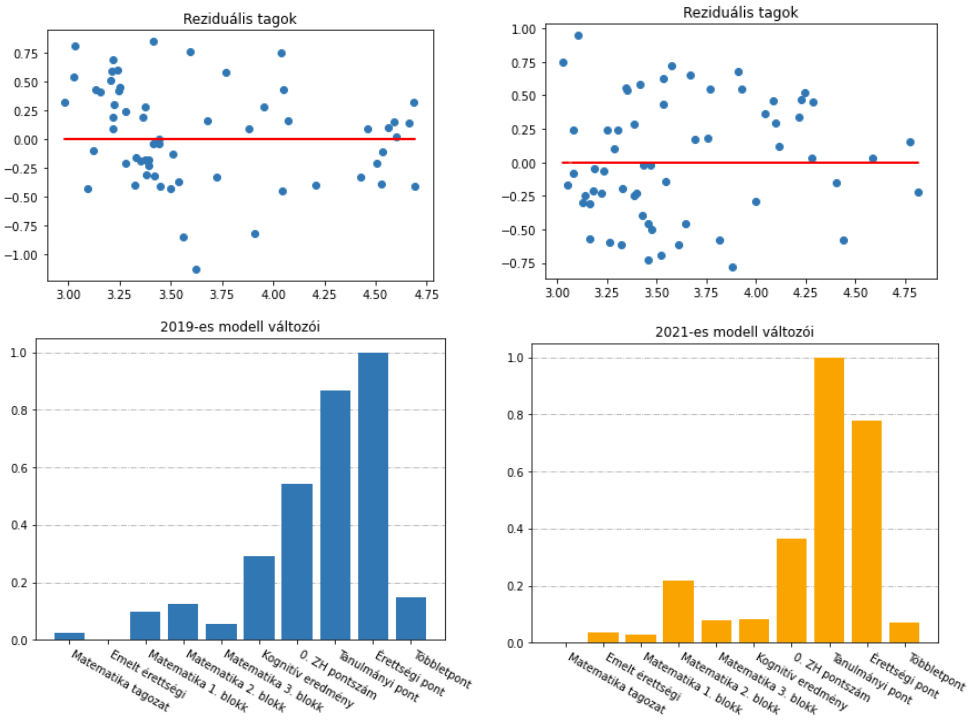
A Gradeint Tree Boosting-gal való előrejelzés eredményei a lineáris regressziónál használt struktúra szerint a 29. ábrán láthatóak.

2019

- R^2: 0.7485
- MAE: 0.3059
- $RMSE$: 0.4253

2021

- R^2: 0.4696
- MAE: 0.3770
- $RMSE$: 0.4407



29. ábra. Gradient Tree Boosting eredménymutatói a két évben

Gradient Tree Boosting-ot használva a 2019-es modell statisztikái drasztikusan jobbak, mint a lineáris regressziónál számított értékek, ugyanakkor a 2021-es modell hatásfoka valamelyest csökkent. A reziduális tagok szórásképe a két évben igencsak eltérő. Ami megjegyzendő, hogy az egyes attribútumok egymáshoz viszonyított prediktív ereje a két évben közel azonos: a két legfontosabb változó a tanulmányi és érettségi pont, viszont a többi változó közül csak a 0. ZH pontszám rendelkezik nem elhanyagolható szignifikanciával.

8. Diszkusszió, következtetések a modellezésről

A matematika érdemjegyek prediktálása során sikeresebben megjósolhatóak a 2021-ben kezdett hallgatók teljesítményei, amely azt jelenti, hogy náluk kisebb az ingadozás és a variancia a bemeneti adatok tekintetében a célváltozóra nézve. Ugyanakkor a legjobb modellekkel is csak 80-85% közötti hatásfok érhető el, amely bár kellően jobb, mint egy véletlen osztályzó által adott 33%-os illetve 50%-os eredmény, de még bizonyosan tovább növelhető. A teljesítmény több adatponttal és több, új bemeneti változó (középiskola földrajzi lokációja; nem; életvitellel kapcsolatos adatok stb.) vizsgálatával mindenképpen javítható lenne, ugyanakkor az új érettségi rendszer 2024-re esedékes bevezetésével érdemes ezek mellett új modellezési struktúrákat is kialakítani. Egy másik megoldás a teljesítmény növelésére egy ensemble modell lenne, amely az egyes kategóriákban legjobb teljesítő, előre meghatározott számú algoritmus eredményét vetné össze és súlyozná az algoritmusok hatásfokával, majd ezek alapján adna egy végső predikciót, ugyanakkor ez több időt és magasabb szintű optimalizációt venne igénybe. A bemeneti változók tekintetében a legfontosabb észrevétel, hogy 2021-ben drasztikusan megnőtt a felvételi pontszám jelentősége a matematika érdemjegyre nézve, viszont a kognitív készségek érdemjegyre gyakorolt hatásaiban csökkentek a különbségek a hallgatók között.

A kumulált átlag predikcióinál látható, hogy a két legjobban domináló tényező mindkét évben a tanulmányi és az érettségi pont, a többi attribútum pedig változó, de többnyire elhanyagolható szignifikanciával bír. Az is észrevehető, hogy egy "black box", nehezebben értelmezhető modell, mint a Gradient Tree Boosting algoritmus, sokkal több információt tud kinyerni az érettségi és tanulmányi pontokból a kumulált átlagra nézve, mint a lineáris regresszió, így további modellezések során célszerű az előbbit használni, esetleg más ensemble módszereket alkalmazni. Ugyanakkor a lineáris regressziót érdemes lehet kibővíteni harmad- vagy negyedfokú polinomiális regresszióvá, vagy kombinálni az osztályozófüggvényt valamilyen trigonometrikus függvénnyel, hisz a lineáris regressziónál kapott két szóráskép hasonló görbét ír le, s így pontosabb eredményekhez juthatunk mindkét évben. Természetesen a teljesítmény több adatponttal és több változóval itt is potenciálisan tovább javítható lenne.

Egy fontos és nem triviális kérdés az, hogy ezeket az eredményeket milyen módon lehet felhasználni. Egyfelől a különböző attribútumok prediktív erejének változását figyelembe véve a célszerű lenne az egyetemek (jelen esetben főként a BME) tanrendébe illetve felvételi rendszerébe beépíteni az észrevételeket. A 2024-es felvételi eljárás módosításának köszönhetően az egyetemek nagyobb szabadságot kapnak a felvételi pontok számítását illetően, így a szabadon kioszthatő 100 pont keretein belül az észrevételek felhasználásra kerülhetnének. Amennyiben a dolgozatban taglalt kutatást kiterjesztenénk több egyetem hallgatóinak vizsgálatára is, akkor akár egyetem-specifikus változtatásokat is meg tudnánk határozni. Egy másik felhasználási mód az lenne, hogy az optimális modellekkel még a szemeszter elején kiszűrjük a nagy valószínűséggel lemorzsolódó hallgatókat, és esetleg felzárkóztató kurzusukat kínáljunk nekik. Ami fontos szempont, hogy ezt mennyire eti-

kus formában lehet megvalósítani, hiszen ha rögtön év elején azzal szembesítjük hallgatót, hogy nagy valószínűséggel lemorzsolódik, az egyes személyeknél demoralizálóan tud hatni, és képes rányomni a bélyegét a hosszútávú egyetemi teljesítményükre. Ezt a problémát részben orvosolhatná az, ha az egyes változók fontosságát és a modellezés, osztályozás folyamatát adott bemeneti paraméterek mellett az egyszerű felhasználó számára is világos, könnyen értelmezhető, esetleg interaktív módon tudnánk vizualizálni. Molontay et al. [11, 12] révén korábban már készültek kutatások, amelyek a magyar felsőoktatási felvételi rendszerben használt bemeneti adatok értelmezhetőségének javítására irányultak a későbbi egyetemi teljesítmény ismeretében. Az általuk vázolt metodológiával összhangban a LIME [30] és SHAP [29] Python könyvtárak megfelelő adatvizualizáció mellett remek megoldást nyújthatnának erre a problémára.

9. Összefoglalás

Munkánkban modern adattudományi eszközökkel vizsgáltuk a 2019-ben és 2021-ben beiratkozott elsőéves BME VBK hallgatók első félév végi teljesítményében és bemeneti adataiban fellépő különbségeket. A dolgozat első felében először különféle szórás- és oszlopdiagramok segítségével vizsgáltuk a két év adataiban előforduló eltéréseket, majd többféle folyamatábrával és klaszterezési kísérlettel folytattuk az elemzést mélyebb kapcsolatokat keresve. Az ábrák ugyan rávilágítottak néhány fontosabb összefüggésre, viszont a részletesebb elemzések szükségességét is megerősítették.

A dolgozat második felében prediktív analitikai módszerekkel vizsgáltuk az egyes bemeneti adatok fontosságának változását első félév végi teljesítménymutatókra nézve, valamint annak a hatásfokát, hogy ezen mutatók értékei mennyire jól előrejelezhetők. A modellezés során az egyes prediktáló algoritmusok mellett főkomponens analízist is alkalmaztunk. Az eredményekből az szűrhető le, hogy a 2021-ben kezdett hallgatók eredményei ugyan romlottak a járvány kitörése előtti szinthez képest, ugyanakkor sok esetben pontosabban előrejelezhető a teljesítményük, valamint a felvételi pontszám jelentősebb prediktáló erővel bír az első féléves matematika jegyre és kumulált átlagra nézve, mint 2019-ben.

Legjobb tudásunk szerint ez az első olyan kutatás, amely ilyen elgondolással és metodológiával vizsgálja a hallgatók bemeneti adatainak megváltozását a pandémia előtti és alatti időszakban, így kutatásunk egyedülállónak mondható. További céljaink közé tartozik a kutatás többféle bemeneti változóval való kibővítése, új modellezési struktúrák és algoritmusok kipróbálása valamint az eredmények egyetem- és hallgatóbarát hasznosítása is.

Hivatkozások

[1] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 1189-1232.

[2] Jovial Niyogisubizo, Lyuchao Liao, Eric Nziyumva, Evariste Murwanashyaka, Pierre Claver Nshimyumukiza. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, Volume 3, 100066, 2022.

[3] Marina Segura, Jorge Mello, Adolfo Hernandez. (2022). Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?. *Mathematics*, Volume 10, 3359, 2022.

[4] Ahajjam Tarik, Haidar Aissa, Farhaoui Yousef. Artificial Intelligence and Machine Learning to Predict Student Performance during the COVID-19. *Procedia Computer Science*, 184:835-840, 2021.

[5] D. K. Dake, D. D. Essel, J. E. Agbodaze. Using Machine Learning to Predict Students' Academic Performance During Covid-19. *2021 International Conference on Computing, Computational Modelling and Applications*, pages 9-15, 2021.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, Volume 12, pages 2825-2830, 2011.

[7] C. Cortes, V. Vapnik. Support-vector network. *Machine Learning*, Volume 20, pages 273–297, 1995.

[8] Roland Molontay. *Lecture notes in Intorduction to Data Science.* Budapest University of Technology and Economics, 2021.

[9] Bolla Marianna, Krámli András. *Statisztikai következtetések elmélete.* Elméleti matematika. Typotex Kft., 2012.

[10] Jonathon Shlens. A Tutorial on Principal Component Analysis. *CoRR*, abs/1404.1100, 2014.

[11] Máté Baranyi, Marcell Nagy, Roland Molontay. Interpretable Deep Learning for University Dropout Prediction. *Proceedings of the 21st Annual Conference on Information Technology Education*, pages 13-19, 2020.

[12] Marcell Nagy, Roland Molontay. Comprehensive analysis of the predictive validity of the university entrance score in Hungary. *Assessment & Evaluation in Higher Education*, 46:8, 1235-1253, 2021.

[13] J. Cavanaugh, S. Jacquemin, C. Junker. A look at student performance during the COVID-19 pandemic. *Quality Assurance in Education*, Vol. ahead-of-print, No. ahead-of-print, 2022. https://doi.org/10.1108/QAE-01-2022-0008

[14] El-Sayed Atlam, Ashraf Ewis, M.M. Abd El-Raouf, Osama Ghoneim, Ibrahim Gad. A new approach in identifying the psychological impact of COVID-19 on university student's academic performance. *Alexandria Engineering Journal*, Volume 61, Issue 7, pages 5223-5233, 2022.

[15] E.M. Onyema, N.C. Eucheria, F.A. Obafemi, S. Sen, F.G. Atonye, A. Sharma, A.O. Alsayed. Impact of coronavirus pandemic on education. *Journal of Education and Practice*, Volume 11, pages 108-121, 2020.

[16] Yamini Chandra. Online education during COVID-19: perception of academic stress and emotional intelligence coping strategies among college students. *Asian Education and Development Studies*, Volume 10, No. 2, pages 229-238, 2020.

[17] Kimkong Heng, Koemhong Sol. Online learning during COVID-19: Keychallenges and suggestions to enhance effectiveness. *Cambodian Journal of Educational Research*, Volume 1, No. 1, pages 3–16, 2021.

[18] Mohammad Alawamleh, Lana Mohannad Al-Twait, Gharam Raafat Al-Saht. The effect of online learning on communication between instructors and students during Covid-19 pandemic. *Asian Education and Development Studies*, Volume 11, No. 2, pages 380-400, 2022.

[19] K. Mukhtar, K. Javed, M. Arooj, A. Sethi. Advantages, Limitations and Recommendations for online learning during COVID-19 pandemic era. *Pakistan Journal of Medical Sciences*, S27–S31, 2021.

[20] Waqas Haider Bangyal, Rukhma Qasim, Najeeb ur Rehman, Zeeshan Ahmad, Hafsa Dar, Laiqa Rukhsar, Zahra Aman, Jamil Ahmad. Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches. *Computational and Mathematical Methods in Medicine*, Volume 2021, Article ID 5514220, 2021. https://doi.org/10.1155/2021/5514220

[21] Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, Raviraj Joshi. Evaluating Deep Learning Approaches for Covid19 Fake News Detection. *CoRR*, abs/2101.04012, 2021. https://arxiv.org/abs/2101.04012

[22] Marianna Isaakidou, Emmanouil Zoulias, Marianna Diomidous. Machine Learning to Identify Fake News for COVID-19. *Stud Health Technol Inform*, Volume 281, pages 108-112, 2021.

[23] Akash Gupta, Amir Gharehgozli. Developing a machine learning framework to determine the spread of COVID-19 in the USA using meteorological, social, and demographic factors. *International Journal of Data Mining, Modelling and Management*, Volume 14, No. 2, pages 89-109, 2022.

[24] Y. Alali,F. Harrou, Y. Sun. A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models. *Sci Rep*, 12, 2467, 2022.

[25] Z. Malki, E.S. Atlam, A. Ewis, G. Dagnew, A.R. Alzighaibi, G. ELmarhomy, M.A. Elhosseini, A.E. Hassanien, I. Gad. ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound. *Neural Comput & Applic*, Volume 33, pages 2929–2948, 2021.

[26] Judit Monostori. The school year 2020-2021 in Hungary during the pandemic - Country report. *Publications Office of the European Union*, Luxembourg, 2021.

[27] Proháczik Ágnes. A tantermi és az on-line oktatás (tanítás és tanulás) összehasonlító elemzése. *Opus et Educatio*, Volume7. No.3, 2020. `http://opuseteducatio.hu/index.php/opusHU/article/view/390/672`

[28] P. Csépe, E. Dinya, P. Balázs, S. M. Hosseini, G. Küzdy, L. Rosivall. Impact of the first wave of COVID-19 pandemic on the Hungarian university students' social and health behaviour. *Z Gesundh Wiss*, 28:1-7, 2021.

[29] Scott Lundberg, Su-In Lee. A Unified Approach to Interpreting Model Predictions *Advances in Neural Information Processing Systems 30*, Curran Associates Inc., pages 4765-4774, 2017.

[30] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. "Why Should {I} Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, pages 1135-1144, 2016.

[31] M. Ester, H. P. Kriegel, J. Sander, X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, AAAI Press, pages 226–231, 1996.