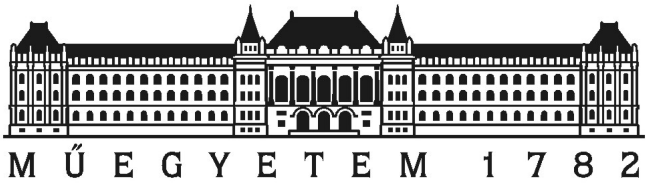


TDK DOLGOZAT

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

....

2. A kognitív tesztről

(ez a rész nem tudom, hogy kell-e)

3. Adatprepació

3.1. Adatok jellemzése, adattisztítás

Az adatokat az EduBase és Neptun rendszerből lekérve nyertük több adattábla formájában. A táblákból a kutatáshoz használt adatok egyrészt olyan bemeneti adatok, amelyeket a szemeszter első hetéig bezárólag érnek el a hallgatók, másrészt olyan teljesítménymutatók, amelyekről csak a félév végén van információnk (utóbbiak prediktálására építünk modelleket a prediktív analitikai részben). Ezek egy része alapvetően rendelkezésünkre állt a táblákból, másokat a meglévő sorokból és oszlopokból megfelelő *feature engineering* segítségével hoztunk létre. A változók pontos 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 kognitív teszt kognitív készségeket mérő részén elért százalékos teljesítmény (0-100-as skálán).
0. ZH pontszám	A BME központi 0. ZH-ján 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.
Kumulált átlag	Az első félév végén 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 kognitív eredményeket tartalmazó adatsor tisztítására volt a legnagyobb szükség, ugyanis az EduBase rendszerben az egyes oszlo-

pokra 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.

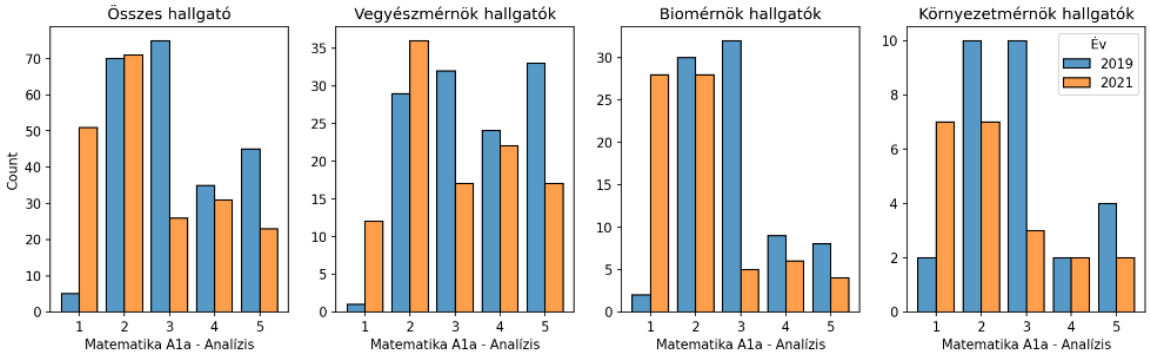
4. Felderítő adatelemzés a két évben

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 a 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 legérdekesebbeket. 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.

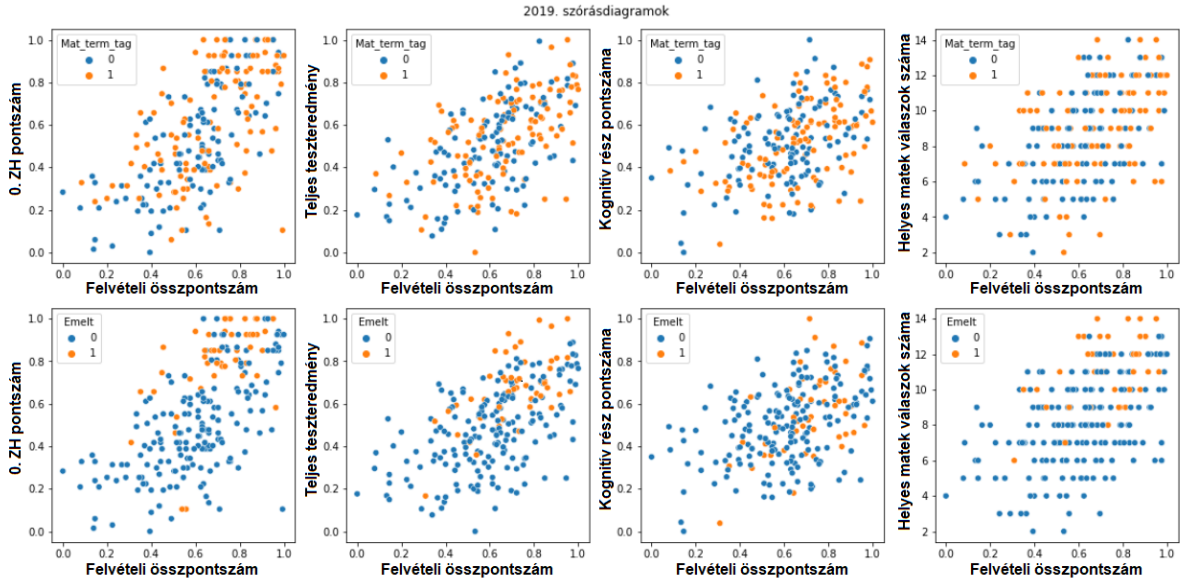
Az 1. ábrán az két évben szerzett Matematika A1a -Analízis jegyek eloszlása látható összesítve és szakokra lebontva.



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 érdmejegyek 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 *seaborn pairplot* függvény előnye, hogy csak az általunk vizsgált változók oszlopait kell megadni, és az összes kombinációt egyszerre számítja ki és ábrázolja egy megfelelő méretű mátrixban. Először tekintsük a *matplotlib figure* alakzatokba összesített ábrákat. Ezeknek az az előnye a *seaborn pairplot*-tal szemben, hogy tetszőleges ábrákat lehet belehelyezni, így például színezéssel bevihető egy harmadik dimenzió a szórásdiagramba. Sőt negyedik dimenziót is lehetne ábrázolni a pontok kinézetének módosításával, de az már nehezen értelmezhető.



2. ábra. 2019. Szórásdiagramok

A 2. ábrán egyrészt a 0. ZH, a teszt összpontszáma és a két külön rész pontszámai vannak ábrázolva pontfelhőként a felvételi összpontszám függvényében, de minden eset 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).

Egyrészt rögtön látszik, hogy minden ábrán a bal felső és a jobb alsó sarokban többnyire nincs pont, ez a pozitív korreláció, azaz összefüggés jele. Másrészt a színezésből arra következtethetünk, hogy az emelt érettségi megléte erősen fölfele húzza a többi eredményt, míg a természettudományi tagozatosok esetében nem egyértelmű a helyzet, drasztikus elkülönülés nincs.

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 lesz némi redundancia a lentebb ismertetett ábrákban.
Dolgozatunkban hat opciót vizsgáltunk a két évben az elérhető adatoknak megfelelően. Mindkét esetben az állapotok a hallgatók valamely eredményét jelentették, például felvételi pontszámot.

- 2019:

    1. matematikai teszt 1.,2. és 3. részeiben helyesen megoldott feladatok száma.

2. kognitív teszt eredmény - *A1* jegy - *A2* jegy.

- 2019 és 2021:

1. tanulmányi hozott pontszám - érettségi pontszám - *A1* jegy.
2. felvételi pontszám - nulladik ZH - kognitív teszt eredmény.
3. felvételi pontszám - kognitív teszt eredmény - *A1* jegy - kumulált átlag.
4. felvételi pontszám - kognitív teszt eredmény - első féléves átlag - kumulált átlag.

Az első két esetben csak a 2019-es évfolyamról volt adat. Az utolsó esetben 2021-ben megegyezett a kutatás idejében az első féléves és a kumulált átlag, így ennél az utóbbi értelemszerűen nem lett még egyszer beletéve.

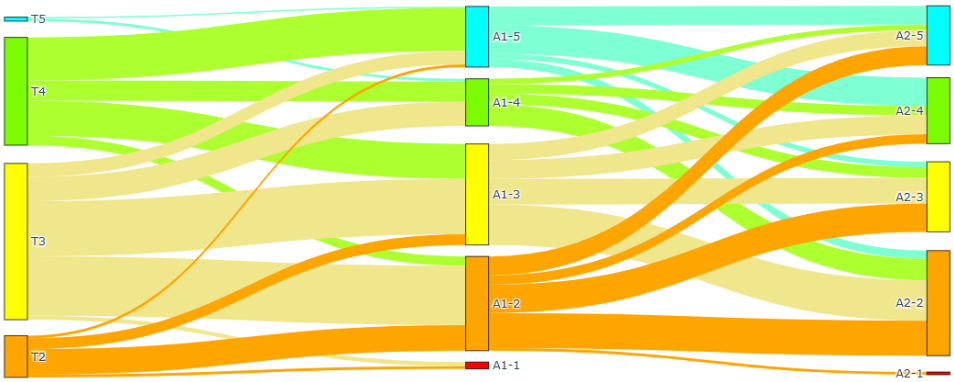
A matematikai teszt részeinek eredményein és az *A1*, *A2* jegyeken kívül a többi változó folytonos volt, így szükség volt ezek diszkretizálására. Öt-öt osztály lett létrehozva minden esetben, de nem feltétlenül került mindegyikbe hallgató. Az átlagok kerekítve lettek, a felvételi, kognitív teszt és nulladik zh pontszámok ekvidisztáns módon lettek felosztva.

A *felvételi pontszám - nulladik ZH - kognitív teszt eredmény* változatban a nulladik zh és a teszt eredményei nullától az elérhető pontszámig lettek felosztva 20%-os lépésközökkel. Továbbá a felvételi pontszám 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.

Ezután *for ciklus* használatával elkészültek a tranzíciós mátrixok. Ezen mátrixok i-edik sorának j-edik eleme azt mutatta, hogy az egyik változó i-edik osztályában hány olyan hallgató volt, aki a másik változó j-edik osztályába került. Ez a folyamok szélességének megadásához volt szükséges.

A *plotly.graph_objects* könyvtár *go.Sankey* függvényével lettek elkészítve egyenként az ábrák. Minden esetben meg kellett adni számozva a kiinduló és beérkező állapotokat, így a középső csoportok mindkét listában szerepeltek. A program index alapján kapcsolta össze a két listából az állapotokat, továbbá kellett egy harmadik lista, amelynek a megfelelő indexű eleme a két állapot közötti folyam mérete. Ezen kívül a folyamokhoz és magukhoz az állapotokhoz is színeket kellett még rendelni, ismét egy-egy listában, figyelve az indexeket.

Végül a kirajzolódott ábrát kellett még kézileg igazítani, 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 nem okoztak gondot, ugyanis a csoport mozgatásával a hozzátartozó sávok automatikusan mozogtak együtt.



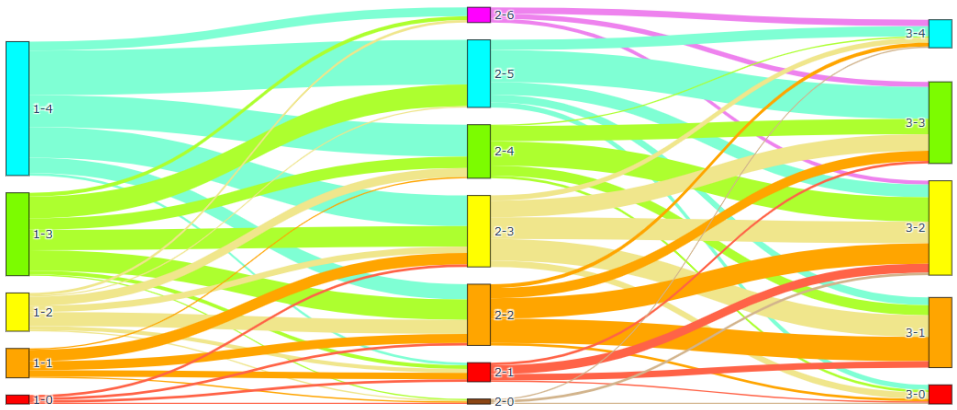
3. ábra. 2019. Kognitív teszt eredmény - *A1* jegy - *A2* jegy

A 2021-es gólyák a kutatás idejében még nem vehették fel az *A2* tárgyat, így értelemszerűen csak a 2019-es évfolyamra készült ez az ábra.

Az 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 sejthető, hogy egy hallgató teljesítménye a teszt során összefügg a további teljesítményével, de viszonylag sok esetben van jelentős romlás, illetve javulás.



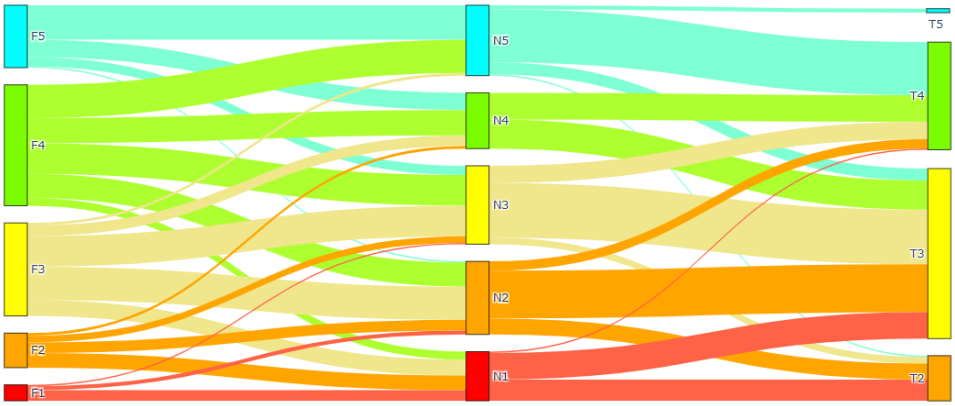
4. ábra. 2019. Matematikai teszt 1. blokk - 2. blokk - 3. blokk helyes válaszok száma

A teszt matematikai részének a komplexitása a blokkonként nőtt, így érdekes volt megvizsgálni, hogyan teljesítettek a hallgatók, és hogyan viszonyul egymáshoz a helyes válaszaik száma a különböző blokkokban.

Rögtön látszik, hogy az első blokkban a hibátlan (négyes) osztály a legnagyobb, és a rosszabb eredmények száma csökken. Ez várható, hiszen itt alapismeretek voltak felmér-

ve. A második és harmadik blokknál teljesen más a kép, a közepes osztályok körülbelül azonosak, a szélsők pedig kicsik, jobban hasonlít a normális eloszlásra.

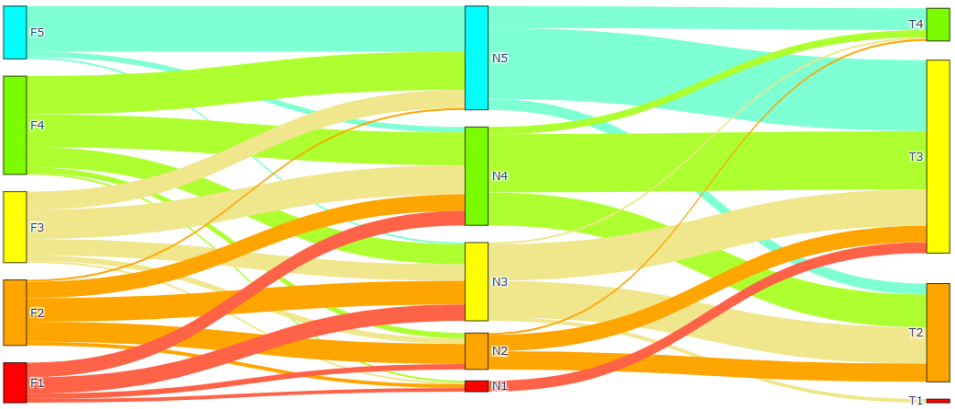
A folyamok egészen változatosak. Az ábra közepén erősen szétoszlanak, de a széleihez közelítve kisebb trendek megfigyelhetők. Például a második blokkban minden kérdésre helyesen válaszolt hallgatók az első blokkban is többnyire szinte csak helyesen válaszoltak.



5. ábra. 2019. Felvételi - 0. zh - kognitív teszt pontszámok

Az 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 kognitív teszt eredménye.

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.

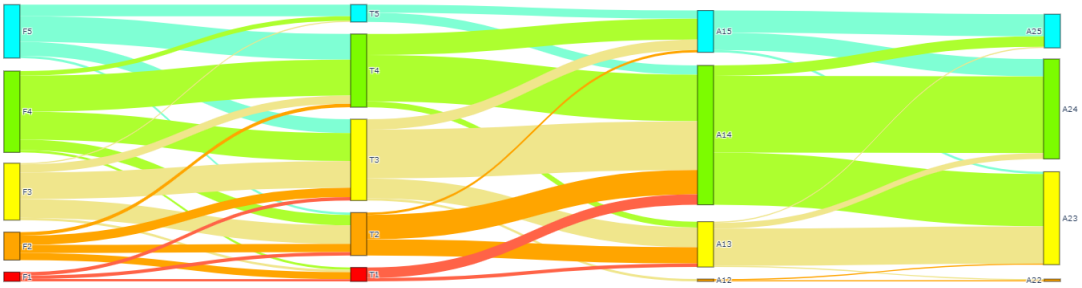


6. ábra. 2021. Felvételi - 0. zh - kognitív teszt pontszámok

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 érdekes, hogy akiknek a legrosszabb lett a 0. zh, a kognitív teszt közepesen ment. Itt is mondható, 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.

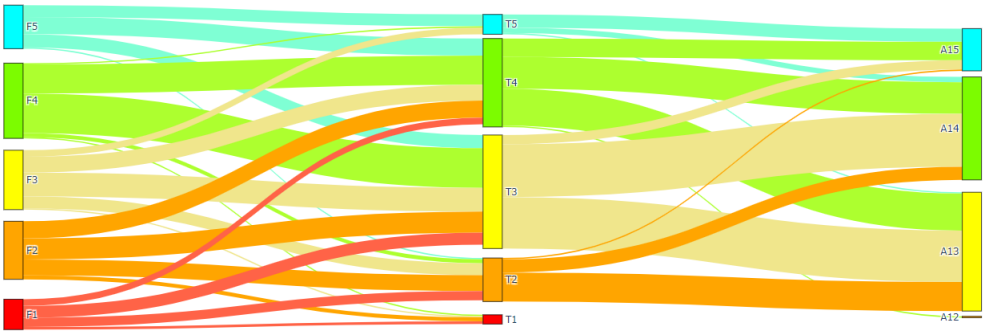


7. ábra. 2019. Felvételi - kognitív teszt pontszám - első féléves átlag - kumulált átlag

Ezen az 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.

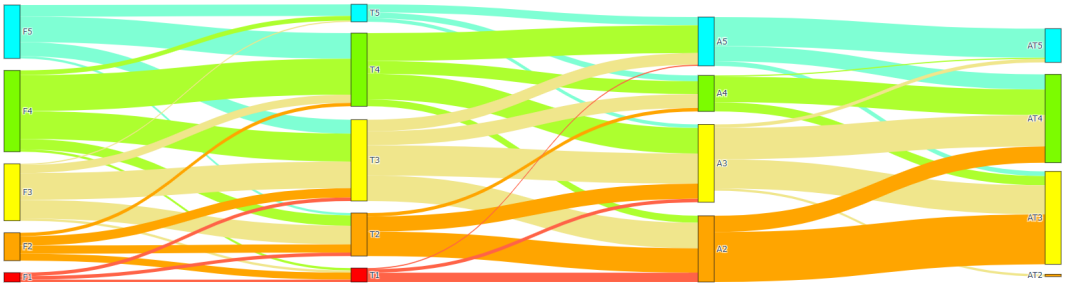
Két előnye volt az értékkészlet szerinti osztályozásnak: gyorsabb volt (egy függvényre volt csak szükség), illetve így az eredmények egymáshoz képest való relatív eloszlása jobban látszik. Míg a régebbi ábrán nagyon nagy a hármas osztály, mert a legtöbben 40-60% körüli eredményeket értek el, itt az látszik, 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 az 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.



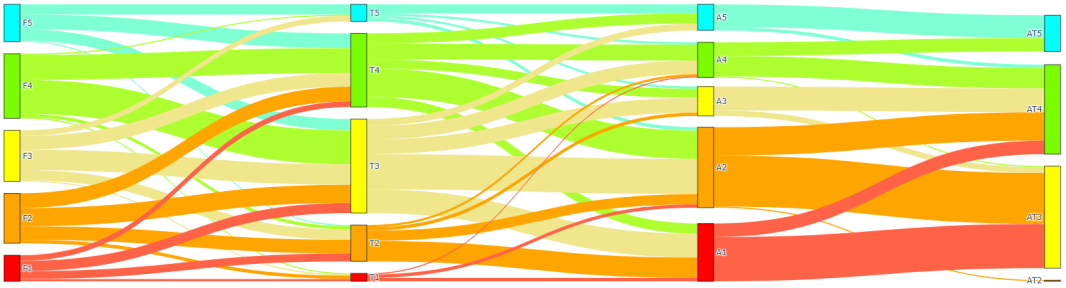
8. ábra. 2021. Felvételi - kognitív teszt pontszám - első féléves átlag

A kutatás során ez az évfolyam még az első évében volt, így csak első féléves átlagokról volt adatunk. Ettől eltekintve 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.



9. ábra. 2019. Felvételi - kognitív teszt pontszám - *A1* jegy - kumulált átlag

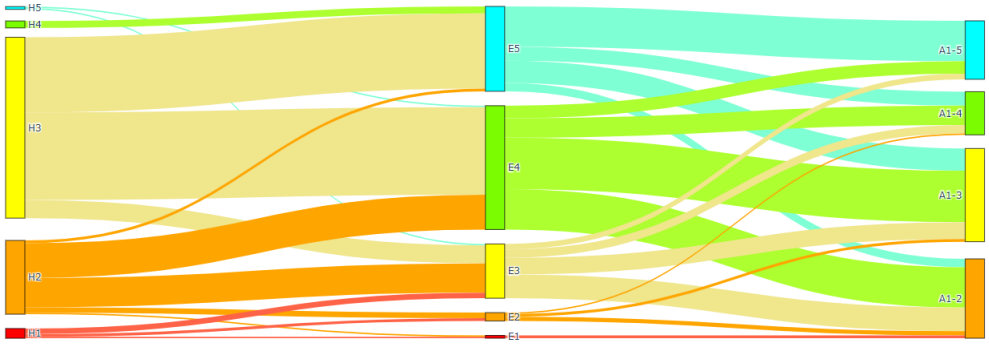
Az első két oszlop viszonyát már láttuk, az újdonság az *A1* jegy az első féléves átlag helyett. Ettől függetlenül a tendenciák változatlanok, azok szereztek jobb, illetve rosszabb jegyet, akik rendszerint jobban, illetve rosszabban írták meg a tesztet. A trendet továbbra erősíti a jegyek és átlagok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



10. ábra. 2021. Felvételi - kognitív teszt pontszám - *A1* jegy - első féléves átlag

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 is, 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. Érdekes, hogy az egyest és kettest szerzett hallgatók mégis nagyrészt hármas, sőt nem kevesen négyes átlagot szereztek. Ettől függetlenül a szokásos trendek itt is megmutatkoznak.

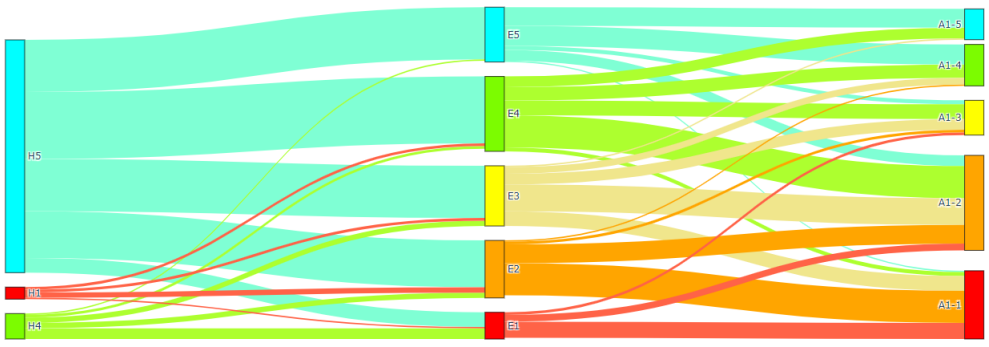


11. ábra. 2019. Hozott és - érettségi pontszám - *A1* jegy

Ahogy később látni fogjuk, a modellek alapján a hozott és érettségi pontszámok erősen befolyásolták a jegyekre vett predikciókat, így utólag elkészült ez és a következő ábra.

Jellegzetes, hogy a hallgatók nagy részének hasonló volt a hozott 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.



12. ábra. 2021. Hozott és - érettségi pontszám - *A1* jegy

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

Ehhez, és a 2019-es ábrához képest az érettségi pontszám viszont egészen egyenletesen oszlott el. Az ábra bal felétől eltekintve, a teljesítmény megmaradására vonatkozó tendencia igaz gyengébben, de ismét jelen van.

4.3. Klaszterezés

EZ MÉG CSAK RÁVEZETÉS

Cél: gép által csoportosítani a hallgatókat több változó alapján, keresni jó magyarázó változókat, hogy egyszerűsíthető legyen a csoportosítás.

Változók: összesített teszteredmény, külön matek/kognitív eredmény, első félév átlaga, kumulált átlag, felvételi pontszám összesítve és részekre bontva. Ezen attribútumok nem redundáns részhalmazai. Az algoritmusok hatékonyabb és gyorsabb működése érdekében a változók értékei normalizálva és skálázva lettek. Három-három táblázat készült a 2019, 2020, 2021 évekre: min-max, illetve kvantilis szerint skálázott, továbbá standardizált értékekkel. Három klaszterező algoritmus lett alkalmazva a legjobb eredmény elérése érdekében: K-közép, DBSCAN és Ward féle hierarchikus klaszterező algoritmus. Ezek futtatása előtt el lett végezve egy felderítő ábrázolás, amely nehézségekre hívta fel a figyelmet. Az adatok kétdimenziós vetületeit ábrázoltuk szórásdiagramokon. A pontfelhők között nem volt olyan, amelyen látványosan el lehetett volna különíteni egyes csoportokat, ami a klaszterezést nehezítette.

---

5. Prediktív analitika

5.1. Modellek és metodológia

Következő lépésként azt vizsgáltuk, hogy a két évben külön a bemeneti adatok alapján mennyire pontosan lehet előrejelezni egyes teljesítménymutatók értéket, 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 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. 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. A legoptimálisabb 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 kutatás során a cél a teszthalmazon való minél hatásosabb előrejelzés mellett az egyes éveknél használt és optimalizált algoritmusok esetén a különböző attribútumok prediktív erejének összehasonlítása volt.

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. (IDE REFERENCIÁK) Mivel azonban viszonylag kevés adat állt a rendelkezésünkre, ezért az érdemjegyek prediktálásánál az adatrekordok esetén a pontos érdemjegy helyett érdemjegy csoportok 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 vonatkozóan illetve szétbontva vegyészmérnök és biomérnök hallgatókra egyaránt (a környezetmérnökökről nem állt rendelkezésre elég adat, így őket a szakonkénti bontásban kihagytuk).

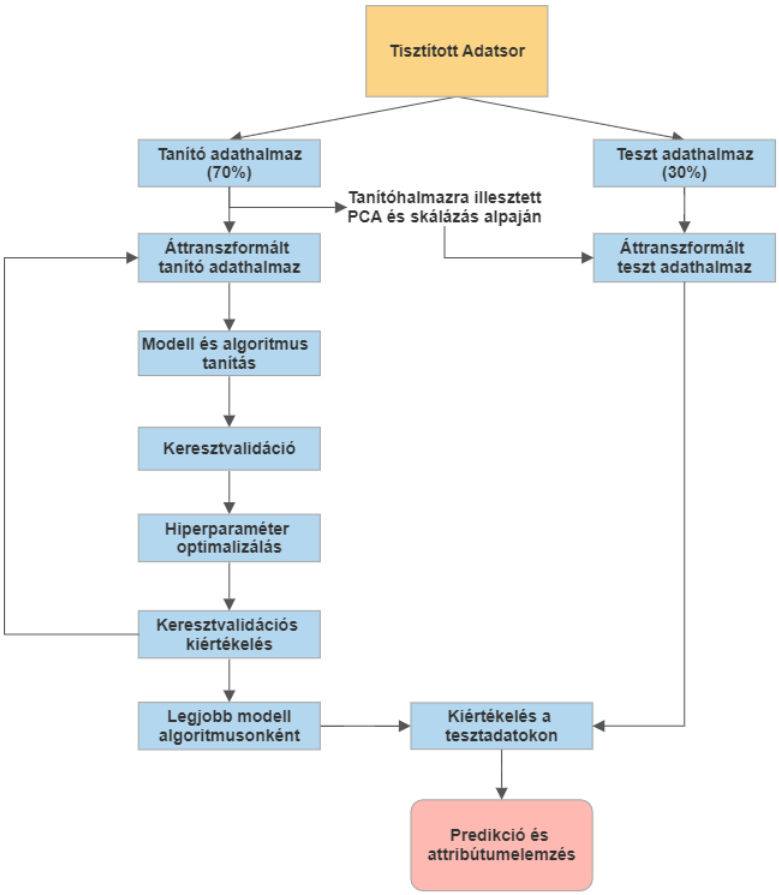
A használandó osztályozó algoritmusok közül a Naive Bayes és Gradient Tree Boosting algoritmusok képesek kezelni a többosztályos feladatokat, a lineáris regresszió azonban folytonos célváltozóérték prediktáláshoz használható, így ott a prediktált értéket kerekítettük a legközelebbi címkeértékhez, a logisztikus regresszió és az SVM pedig 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) is alkalmaztunk. Az eljárás lényege, hogy az adatpontokat egy kisebb dimenziós térre vetítjük le oly módon, hogy a változók közötti varianca minél nagyobb részét tartsák meg, így minimalizálva az információveszteséget. Az új változók, amelyeket főkomponensnek nevezünk, a kisebb dimenziós térben az eredeti változók 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 *jegycsoportok* prediktálására épített modellezési struktúrát 13. 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 skálázásnak vetettük alá, ahol az algoritmusok jobb teljesítőképessége és a változók kiértékelés utáni összehasonlíthatósága végett a különböző attribútumok értékeit a kvantilisek mentén a $[0, 1]$ intervallumba transzformáltuk át, valamint az érdemjegy csoportok prediktálásához a 2 és 3 csoport modelleknél a célváltozó-értékeket rendre 1, 0-ra illetve 3, 2, 1-re módosítottuk.

Valamennyi algoritmus hiperparamétereinek optimális megválasztására 5-szörös keresztvalidációt alkalmaztunk, ahol törekedtünk arra, hogy az adatrekordok címkéjének



13. ábra. A modellezési struktúra sematikus ábrája

eloszlása egyenletes legyen a felosztott részek között. A keresztvalidációnál használt, jóságot mérő metrikának a kiegyensúlyozott pontosságot ('*balanced accuracy*') választottuk, amelynek képlete az alábbi:

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

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 tesztthalmazon visszamértünk, valamint kiértékeltük az ily módon választott modellek esetén az egyes változók fontosságát is.

A *kumulált átlag* predikciójánál a tisztított adathalmazt ugyanúgy 70-30 arányban osztottuk fel. Adattranszformálásra csak skálázást alkalmaztunk, PCA-t nem, majd valamennyi osztályozó hiperparamétereinek optimalizálása hasonlóan 5-szörös keresztvalidációval történt. Az illesztett legjobb modelleknél mindkét évben feljegyeztük a reziduális tagok szórásdriagramját, az egyes változók prediktív erejét, valamint az alábbi statisztikák értékét:

- R^2: A modell hatásfokát mérő mutató, értéke $[-\infty, 1]$ közötti:

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

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

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

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

- $RMSE$: Gyököt vont átlagos négyzetes eltérés:

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

5.2. Osztályozó algoritmusok és optimalizálásuk

5.2.1. 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 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 implementálása során a legkisebb négyzetek elvét alkalmazza, amellyel az optimális megoldás analitikusan elérhető, ezért nem történt hiperparaméteroptimalizáció.

5.2.2. 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.

5.2.3. Gradient Tree Boosting

A Gradient Boosting algoritmus egy Ensemble típusú osztályozó, amelyek lényege, hogy sok gyenge teljesítményű prediktor (*"weak learner"*) eredményét felhasználva hoz egy erős predikciót. Gradient Tree Boosting esetén a gyenge prediktorok *döntési fák*, amelyek lefelé irányított, legtöbbször bináris fák, továbbá minden belső csúcsában egy attribútumra vonatkozó feltétel szerepel, a levelei pedig valamilyen célváltozóértékkel 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)

5.2.4. 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[1] függvény bemeneteként, amelyet a függvény leképez a $(0, 1)$ intervallumra,

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

és amennyiben az output 0.5-nél nagyobb, úgy a pozitív osztályba soroljuk az adott rekordot, különben a negatívba.

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

5.2.5. SVM

Az SVM (Support Vector Machine) algoritmus a logisztikus regresszióhoz hasonlóan egy lineárisan szeparáló hipersíkot akar meghatározni. Lényege, hogy különböző magfüggvénye segítségével az adatrekordokat egy magasabb dimenziós térbe képzi le, ahol olyan szeparáló hipersíkot keres, amely maximalizálja a vele párhuzamos hipersíkok által meghatározot olyan térrészt, amely adatpontot nem tartalmaz. A cél a megfelelő magfüggvény és a legoptimálisabb 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, RBF)

6. 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 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.

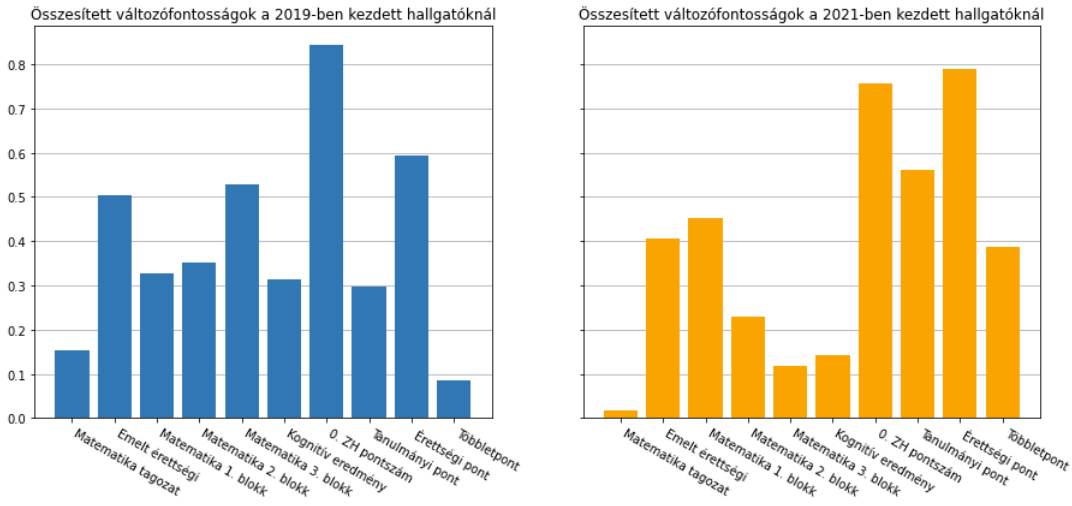
A 3. táblázat a 2021-es adatokon optimalizált algoritmusok eredményét szemlélteti az előző ábráéval megegyező metodológia szerint. Ezen adathalmazon az osztályozók teljesítménye jobbnak nevezhető, mint a 2019-es adatsoron, átlagosan a legeredményesebb algoritmusnak a Naive Bayes nevezhető, amely a szakonkénti 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.

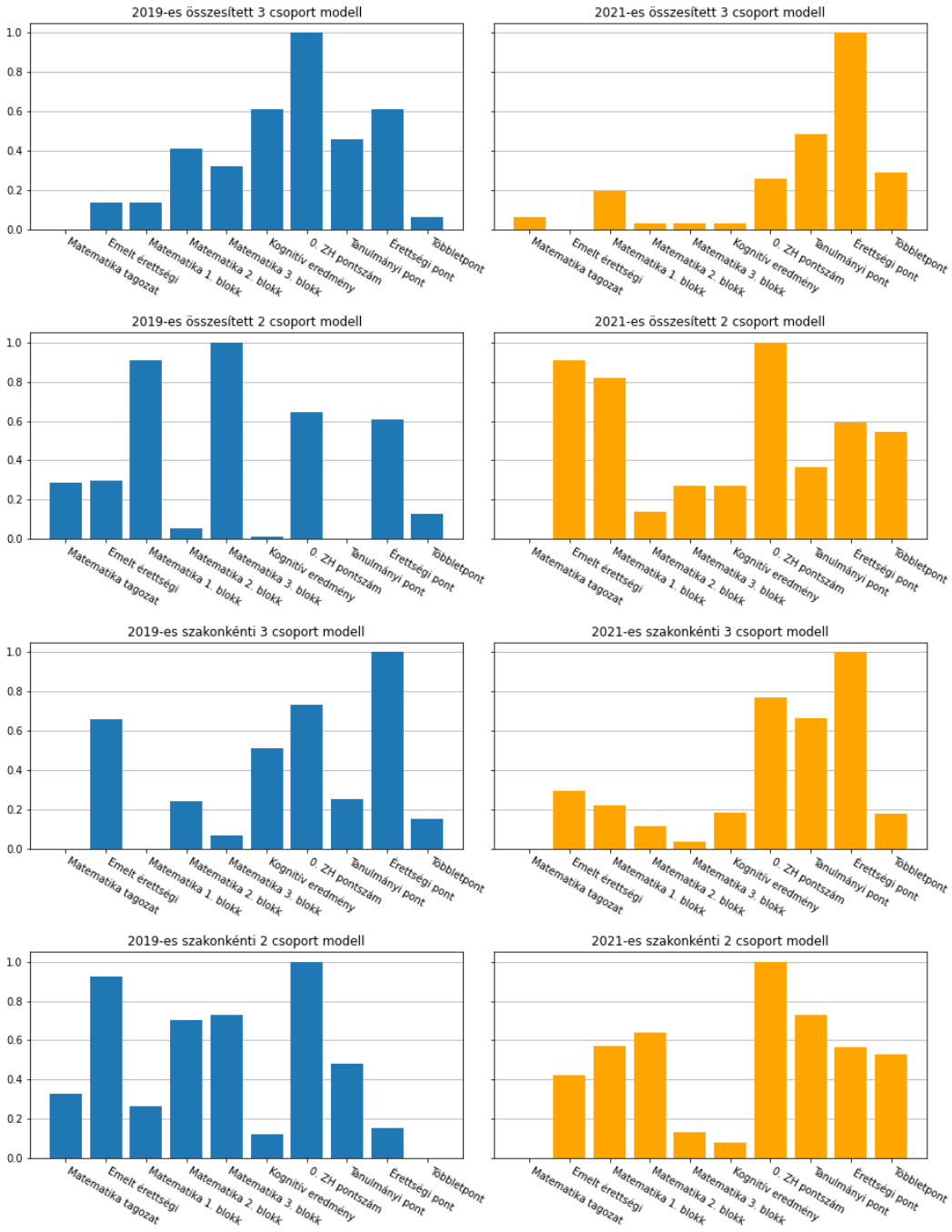
			Osztályozó algoritmusok				
			Grad. Boos.	Naive B.	Log. reg.	SVM	Lin. reg.
3 csoport	Összesített	2 PC	54.24	**80.39**	76.29	78.81	60.92
		4 PC	54.89	72.05	72.13	75.79	70.40
		6 PC	66.24	76.44	70.04	76.94	70.62
		8 PC	63.00	75.29	68.18	73.06	62.07
	Szakonként	2 PC	64.84	81.89	59.23	64.77	49.35
		4 PC	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 14. ábra, míg az egyes modellekhez tartozó attribútumszignifikanciákat a 15. ábra szemlélteti.



14. ábra. Az változók összesített fontossága a két évben

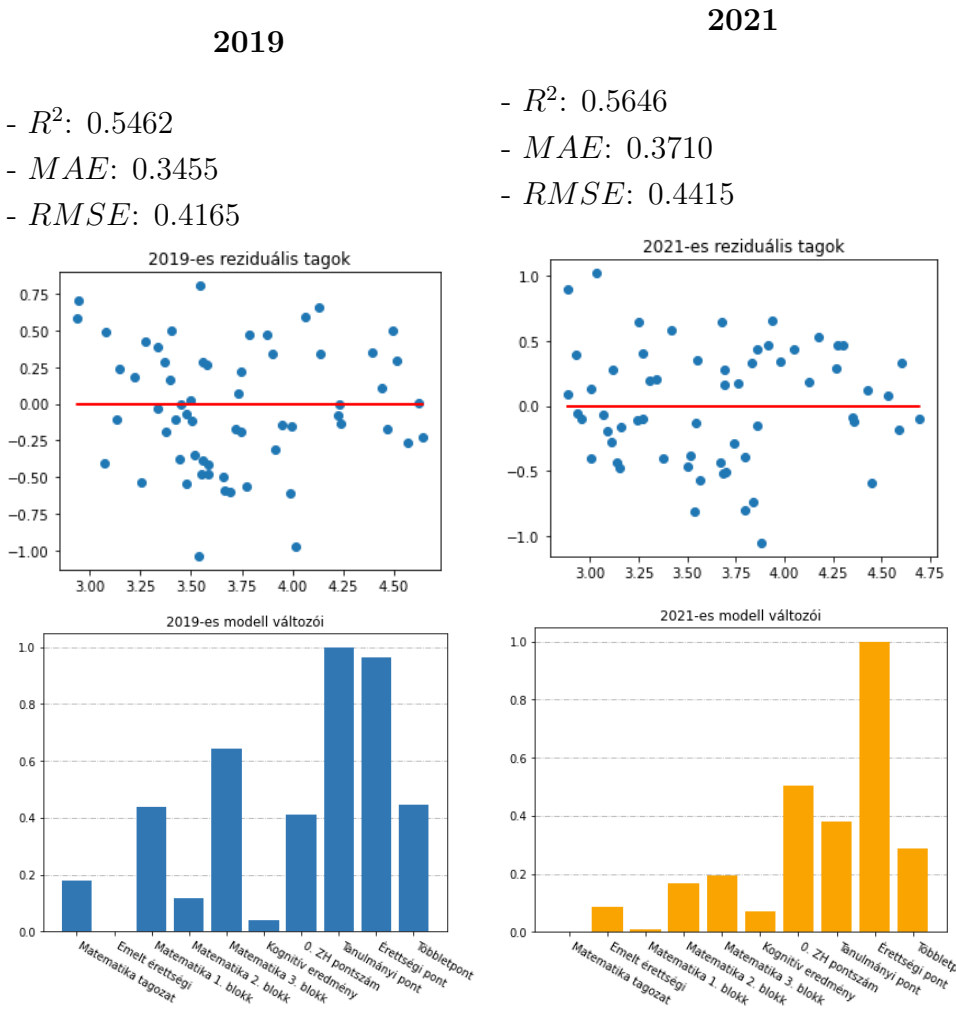


15. ábra. Az attribútumok prediktív ereje a legjobban teljesítő algoritmusoknál

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 kognitív 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. Összevetve a felderítő adatelemzéssel, ezen eredmények arra engednek minket következtetni, hogy ...

A **kumulált átlag** lineáris regresszióval való prediktálásának eredményei (statisztikák, reziduális szórásdiagramok és az egyes változók prediktív ereje egymáshoz viszonyítva) a 17. ábrán láthatóak.



16. 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, amelyek 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 kognitív teszt matematika részén 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.

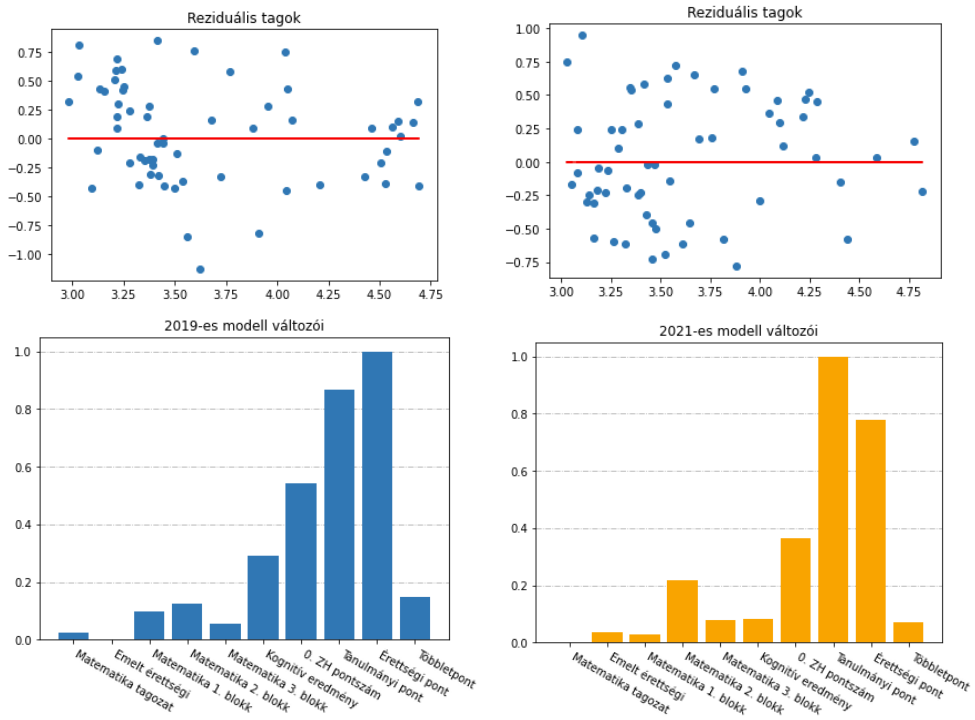
A Gradeint Tree Boosting-gal való előrejelzés eredményei a lineáris regressziónál használt struktúra szerint a 17. ábrán láthatóak.

2019

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

2021

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



17. ábra. Gradient Tree Boosting eredménymutatói a két évben

GBT-t alkalmazva 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.

7. Diszkusszió, következtetések

A matematika érdemjegyek predktálásánál látható, hogy legjobb esetben is csak 80-85% közötti eredmény érhető el, amely már jelentősen jobb, mint egy véletlen osztályzó által adott 33%-os illetve 50%-os teljesítmény, de még nem elegendően pontos. Ezen eredmények 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.

8. Összefoglalás

Ebben a dolgozatban adattudományi módszerekkel vizsgáltuk az elsőéves BME VBK hallgatók teljesítményét bemeneti adatok tekintetében.

Hivatkozások