**Önálló kutatási feladat, összegző dokumentáció**

**BME VBK hallgatók különböző felmérési eredményeinek vizsgálata, és esetleges lemorzsolódás prediktálása**

**Köller Donát, Vlaszov Artúr, Szilágyi Brigitta**

**Kérdésfelvetés:** Mindig is kulcsfontosságú feladat volt az egyetemeken a lemorzsolódás megelőzése, vagy mértékének csökkentése, hisz egy jó képzés egyik ismérve, hogy a felvett hallgatók közül minél többen végezzék el sikeresen. Az egyetem egy teljesen új környezet a középiskolához képest, megváltoznak az elvárások, a leadott tananyag mennyisége és komplexitása, a számonkérések formátuma, a felelősség. Sokaknak nehéz ehhez gyorsan alkalmazkodni, mentalitást váltani. Emiatt az első évben jelentős nehézségeik akadhatnak, amelyek akár bukáshoz és kiábránduláshoz vezethetnek. Ezért lehetőleg minél korábban, az elérhető információk alapján szeretnénk behatárolni, hogy melyek azok a hallgatók, akik esetleg segítségre szorulhatnak egyetemi tanulmányaik elkezdése során, ugyanakkor azokat a hallgatókat is érdemes lehet már korán megtalálni, akik vélhetően kiválóan fognak teljesíteni az új környezethez képest is. Tehát célunk ilyen hallgatók megtalálása, illetve egyúttal a teljesítményértékelések eredményeiből hasznos információkat, esetleg összefüggéseket leszűrni.

Adattudományi szempontból ez egy kiegyensúlyozatlan, többosztályos osztályozási feladat, ahol rendelkezésünkre áll 2019., 2020. és 2021-ből valamennyi adat. Általánosságban a felvételi pontszámok részpontszámokra bontva, a nulladik zárthelyi eredményei, illetve a legfontosabb, a matematikai és kognitív képességeket mérő tesztek eredményei. A 2019-es adatok esetében rendelkezésünkre álltak az A1 és A2 tárgyból elért jegyek is. Az A1 tárgy jegyének prediktálása volt a cél, ennek megfelelően építettünk és futtattunk modelleket. A másik két év esetében leginkább ábrázolás módján vizsgáltuk a különböző eredmények között fennálló összefüggéseket.

A rendelkezésünkre álló, és általunk talált cikkek körében általában nem pont ilyen problémafeltevés szerepelt. Alapvetően a tartós tudás megszerzésének a módszere volt fókuszban. A 21. században a digitális technológiák fejlődésével, illetve a változó, dinamikus munkakörnyezetben szükség van gyors és effektív módszerek használatára az oktatásban, melyeknek a célja, hogy a hallgatók aktuális, gyakorlatban is széles körben alkalmazható, de leginkább tartós tudásra tehessenek szert. Ehhez az utóbbi idők oktatási rendszerei kezdenek elmaradni, így például a BME-n be lett vezetve a mérnöki szakokon egy új oktatási forma, amelyet a “Bodnár, G., Berezvai, Sz., Verasztó, Zs., Szilágyi, B., 2016 Hatékony új módszerek alkalmazása a BME reguláris matematika oktatásában.” cikkben lehet megismerni. Röviden, létre lett hozva egy online, felhő alapú oktatási felület (EduBase), melyen elérhetőek alapvető és kiegészítő tananyagok, mintafeladatok és más segédletek, illetve magába foglal egy online számonkérési rendszert. Emellett meg lettek szervezve felzárkóztató, illetve tehetséggondozó foglalkozások. Ahogy a cikkek, úgy a mi kutatásunk központjában is a matematikai, illetve kognitív képességeket mérő teszt állt. A teszt a következő módon épül fel. Az első részben 14 matematikai feladat van, három szintre osztva. Minden feladat feleletválasztós, négyből egy jó válaszlehetőséggel. Az első szint egyszerű példákat tartalmaz, amelyek az alapvető matematikai tudást mérik, például hatványozást, logaritmikus azonosságokat, kicsit komplexebb törtes számolásokat. A második szinten egy fokkal nehezebb feladatok szerepeltek, melyek hasonlítanak a középiskolában is látott szöveges példákra, ezek még nem igényelnek komoly tudást. A harmadik szint már szokatlan példákat tartalmaz, amelyekhez nem csak a megfelelő szintű tudás szükséges, de komolyabb probléma megoldási és gondolkodási képességek. A második részben nyelvi feladatok vannak, amelyek valamilyen szinten a hallgatók kreativitását és nyelvi képességeit vélik mérni. Míg a tanulmányok többségében az új módszerek hatékonyágát mérték, alapvetően azok sikerességét alátámasztva, egy cikket ki lehet emelni, ugyanis annak a célkitűzése áll legközelebb a miénkhez (Novel prediction test for freshmen at BME, Faculty of Chemical Technology and Biotechnology Brigitta Szilágyi, Gábor Hornyánszky, Szabolcs Berezvai). Abban különbözik viszont, hogy kicsit más felépítésű felmérés alapján vizsgálódik. Összegezve, a cikk tanulsága, hogy egy hallgatónak ezen a teszten elért eredménye összefüggésben van a további sikerességével. Be lett mutatva, hogy azon hallgatók, akik jó eredményeket értek el az egyes részeken, utána többségben jó eredménnyel zárták a tanulmányaikat is (konkréten a matematikai tárgyakon szerzett jegyeket mérték). Ugyanakkor rávilágítottak például az emelt szintű érettségi hatására is, jóval magasabb pontokat értek el általánosságban az emeltesek, illetve később is a sikeresebb hallgatók csoportjába kerültek. Ezen aspektusokat mi is megvizsgáltuk, és hasonló eredményekre jutottunk. Néhányat ebben a dokumentumban is szemléltetünk majd az elkészített ábráink segítségével.

**Adathalmazok jellemzése, előfeldolgozási lépések**

Az adatokat Szilágyi Brigitta Tanárnő bocsátotta a rendelkezésünkre. A 2019, 2020 és 2021-ben kezdett hallgatókról is voltak adataink, azonban mi a vizsgálódásaink nagy részét a 2019-es adatokra szorítkozva végeztük el, mert csak azok voltak kellően részletesek (a feladat kezdeti fázisában azonban a többi évhez kötődő adattáblán is hajtottunk végre adattisztítási lépéseket, csak később ezt nem tudtuk teljes mértékben hasznosítani). 2019-hez kapcsolódóan több adattáblánk volt, az egyikben a hallgatók felvételi eredménye szerepelt, a másikban a 0.ZH pontszámuk, egy harmadikban a kognitív teszten elért eredményeik, illetve egy negyedikben a Matematika A1a és A2c osztályzataik, illetve még néhány, ahonnan 1-1 adatot kellett kinyernünk. Ahhoz, hogy ezeken dolgozni tudjunk, nyilvánvalóan össze kellett fűznünk őket a hallgatók neve és Neptun kódja alapján. Azonban ez előtt célszerű volt az egyes táblákat egyenként megvizsgálni és alakítani, hogy minél több irreleváns oszlopot ki tudjunk szűrni, így csökkentve a keletkező összefűzött adattábla méretét.

A kognitív eredményeket tartalmazó adattábla részletesen tartalmazott információt egyrészt minden hallgatóról (hova valósi, emelt érettségit tett-e matematikából, reál tagozatos volt-e, milyen szakra és tankörbe jár), másrészt a hallgató teszteredményéről is (mennyi idő alatt töltötte ki a tesztet, mely kérdéseket válaszolta meg jól, milyen lett a nyelvi és matekos teljesítménye), illetve tartalmazott néhány, a teszthez kapcsolódó egyéb információt is (például a teszthez használt edubase jelszó, felhasználónév). Természetesen nekünk ennyi adat nem kell, úgyhogy ebből az adattáblából jó pár irreleváns oszlopot ki kellett szűrnünk. Amelyeket meghagytunk, azok az alábbiak: a hallgató neve (ez már elég volt, hogy csak ez alapján fűzzük össze a táblákat) és Neptun kódja; emelt érettségit tett-e matematikából; reál/matematika tagozatos volt-e; szak és tankör; az elért pont és százalékos teljesítmény a nyelvi és matekos részben, valamint összességében. Problémát jelentett még, hogy a ’Szak’ mezőben mindenki másképp írta be azt, hogy melyik szakon tanul, így ezt szabványosítani kellett, ha később szakok szerint akartunk vizsgálódni. Erre a feladatra külön Python kódot írtunk, és ha volt olyan mező, ahol nem tudtuk eldönteni, hogy mi lenne az oda tartozó érték (például mert ’VBK’ volt odaírva), arra bevezettünk egy globális ’UNKNOWN’ változóértéket, azonban szerencsére ilyenből kevés volt. Kicsit még tisztítani kellett a ’Tankör’ értékeken is, de mivel ilyenből kevés volt, ezt manuálisan is meg tudtuk tenni.

A 0. ZH eredményeket tartalmazó tábla szerencsére ennél jóval kisebb volt, csak a hallgató nevét, Neptun kódját, képzés nevét, illetve kódját, felvétel évét, valamint a ZH eredményt tartalmazta. Ebből értelemszerűen csak a névre és az eredményre volt szükségünk, a többi elhagyható.

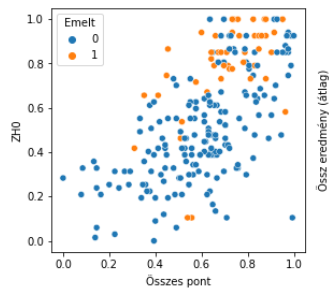
A felvételi pontszámokat bontva (hozott pont, érettségi, többletpont) tartalmazó tábla hallgatók nevén és születési dátumán kívül tartalmazott még pár, a felvételi eljáráshoz és felvételi döntéshez kapcsolódó adatot, illetve a ponthatárt. Ebből a táblából csak a név és a pontokat tartalmazó oszlopok kellettek, a többit elvethettük.

A Matematika A1a és A2c jegyeket tartalmazó táblával már több dolgunk volt, mint az előző kettő esetben. Először is minden személyhez több rekord tartozhatott, legalább egy az A1a jegyhez, lehetett még egy az A2c jegyhez (nyilván csak azoknak, akik elvégezték az A1a-t, és ott maradtak az egyetemen), illetve, ha egy korábbi, nem a végleges jegyet eredményező vizsgán egy hallgató megbukott, akkor ahhoz is tartozott egy rekord. Az attribútumok között a hallgató nevén, Neptun kódján és az osztályzatán kívül szerepelt még a felvétel éve, képzés neve, kódja, státusz ID (Aktív, elbocsátott stb.), Pénzügyi státusz (állami/önköltséges), a tantárgy neve, kódja, kreditértéke, jegy típusa, bejegyzés dátuma, illetve, hogy elismert és hogy érvényes-e az adott jegy. Ezekből az adatokból nekünk csak a hallgató nevére és jegyértékeire volt szükségünk, ráadásul olyan formában, hogy minden sor egy hallgatóhoz tartozzon, és az oszlopok a tantárgyakból szerzett jegyeket tartalmazzák. Ehhez először Pythonban kiszűrtük az irreleváns oszlopokat, majd ’crosstab’-eléssel a kívánt formára hoztuk az adatokat, ahol még ügyelni kellett arra, hogy a korábbi vizsgajegyek ne kerüljenek bele, tehát minden hallgatóhoz tárgyanként csak egy jegy tartozzon. Ezen kívül még, mivel az érdemjegyek szövegesen voltak megadva’, azokat számszerűvé alakítottuk, hogy majd a későbbiekben könnyebb legyen velük dolgozni.

Egy külön adattábla tartalmazta még a kognitív eredményeknél a matekos eredményt blokkokra lebontva, amely valójában az elsőként tekintett adattáblának volt egy egyszerűsített, kevesebb attribútummal bíró változata. Ebből az adattáblából csak a hallgatók nevére, illetve a blokkonkénti teljesítményre volt szükségünk, a többit elhagytuk.

Így már rendelkezésünkre állt az összes tábla, egyenként tisztítva, és már csak az összefűzés volt hátra, amit R-ben könnyen meg tudtunk tenni, valamint még a végén rendeztük az oszlopok sorrendjét, hogy az adathalmaz logikus szerkezetű legyen. Fontos megjegyezni azonban, hogy nem minden elsőéves írt abban az évben kognitív tesztet, így összefűzés során (ahol valójában ’inner-join’-oltunk) kevesebb sorunk lett, mint ahányan abban az évben a BME VBK karára felvételt nyertek.

A legvégén az így kapott adathalmaz 231 rekorddal és 21 oszloppal rendelkezett, amelyek között még esetlegesen szűrtünk különböző algoritmusok használata során.

Ezek után indult a felderítő elemzésünk, mely során kaptunk egy röpke képet a trendekről, kirajzolódott néhány összefüggés, megéreztük, milyen irányban kell mozognunk. Első sorban a Tableau-hoz fordultunk, ugyanis elég egyszerű eszköztára van, és gyorsan tudtunk ábrákat generálni. Ezek egy mellékelt fájlban tekinthetőek meg. Első sorban szórásdiagramokat készítettünk és szemrevételeztünk. (Alapvetően a fájlbeli sorrendben haladok, de igyekszem érthetően megfogalmazni, melyik ábráról lesz szó.) Az első ábrán a 2019. és 2021-es, matematikai és kognitív részben elért eredmények szórásdiagramjait generáltuk egy síkon, és éveknek megfelelően színeztük. Alapvetően hasonló alakot öltenek a pontfelhők, de látható, hogy a 2021-es adatok enyhén balra és lefele tartanak, ami arra utal, hogy gyengébbek, illetve ránézésre kicsit jobban szórnak. A következő ábrán mind a három évnek generáltuk ugyanezen pontfelhőjét, csak most egymás mellé tettük, és a szakok alapján színeztük (UNKNOWN opciót kivéve). Ezeken például az figyelhető meg, hogy 2020-ban matematikából enyhén jobb eredmények születtek a másik két évhez képest, ugyanis sokkal kevesebb rekord került az ábra bal oldalára és aljára. A színezések alapján, ránézésre a vegyészmérnökök eredményei voltak általánosságban jobbak, de erősebben is szórtak. A környezetmérnökök közül pedig viszonylag kevesen értek el kimagasló eredményt a matematikai részen. Végül a harmadik scatterploton azt néztük meg, hogy a szokásos pontfehő hogyan oszlik meg természettudományi tagozat alapján. Sajnos nem túl informatív, körülbelül ugyanúgy oszlanak el az adatpontok, így nyilvánvaló összefüggés a tagozatra járás és a teszteredmények között nincs. Két ábra van még a fájlban, de ezek nincsenek kidolgozva, így ezeket nem vizsgáljuk. A továbbiakban pythonban is készítettünk ábrákat, ugyanis lehetővé teszi hasonló grafikonok gyors generálását, és egymás mellett való ábrázolását nagy alakzatokban. Ezek is mellékelve lesznek, nem néhányat beillesztünk a szövegrész alá. Tekintsük először a 2019-es 2x4-es ábrát. Mindegyik részábrán az x tengelyen az összegzett felvételi pontszám szerepel, a függőleges tengelyen pedig rendre a 0. zh, a teszten elért összesített, majd csak kognitív pontszám, végül a helyesen megválaszolt matek feladatok száma szerepel. Rögtön kiemelhető a trend, akárcsak a fentebb említett cikkben, hogy akik emelt érettségit írtak, azok jelentősen jobb eredményeket értek el a kognitív részt kivéve, ott is jobban szerepeltek, de nem akkora mértékben. A tagozat viszont továbbra sem garantált szebb eredményeket, ugyanis a pontfelhők itt is hasonlóak a két esetben. A 2020-as adatokban rendelkezésünkre álltak felzárkóztató tanfolyam, illetve tehetséggondozó képzéseken való részvételről adatok, de nem volt nulladik zhról adat, így ezt 3x4-es formátumban vizsgáltuk, ahol az oszlopok a különböző bináris attribútumokra vonatkoztak, a sorok pedig a függőleges tengelyen osztozkodtak. A tagozat továbbra sem hordozott látványos változást, az emelt szintű érettségi hatása kicsit gyengült, de többnyire fennmaradt, viszont az utolsó két oszlop szemet szúr. Maximálisan egyértelmű a különbség a két csoport között, igaz itt is a kognitív esetben mosódik összébb. Ennek az az oka, hogy erre a részre nem nagyon lehet felkészülni. A 2021-es adatok alapján az első alakzathoz hasonló struktúrát építettünk fel. A trendek is egészen hasonlóak, bár nem annyira rikítóak, sőt néhány érdekesség is előfordult. Például a tagozatosok kognitív részben és összpontszámban is az ábra alja felé többen vannak, mint a nem tagozatosok. Említésre méltó, hogy az eddigi 3 alakzatban mindenütt fennáll egy enyhe trend, miszerint a sikeresebb felvételizők viszonylag jobban teljesítettek a különböző felméréseken, de meglepően nagy a szórás mindkét irányban, tehát a korreláció pozitív, de nem számíthatunk nagy értékre. Végül, de nem utolsó sorban tekintsük a legnagyobb alakzatot, a 4x4-es ábrát. Itt visszatérünk a 2019-es adatokhoz, és a méretet az indokolja, hogy a tárgyak eredményének függvényében figyeljük meg a többi mutatót, így 8-8 ábra lesz tárgyanként, amiből 4-4 emelt érettségi és tagozat szerint lesz színezve. Sajnos a kétdimenziós pontfelhőkön nem látszik tökéletesen a pontok sűrűsége, csak esetleg minimálisan aszerint, hogy mennyire élénk vagy halvány az ábra adott pontjában a jel, így nem kapunk teljes képet az eloszlásokról. Ugyanakkor az első sor alapján továbbra bizonyodik, hogy a tagozat nincs összefüggésben további eredményekkel. A második sorban az érettségis trend megismétlődik, illetve láthatjuk, hogy jelentősen több a kék (nem emeltes) adatpont a rosszabb jegyek felett. A 3. és 4. sorok folytatják ezeket az irányokat, de figyelemben kell tartani, hogy nem mindenki vette fel az A2 tárgyat, így azoknak 0 az első koordinátája. Összességében mindegyik jegy erőteljesen szór, de itt is látszik egy kissé javuló tendencia. Továbbá, készültek sankey diagramok is, amelyek a bemutatón lesznek részletesebben bemutatva, mellékelve vannak ők is. Róluk is leolvashatóak a megfelelő tendenciák, de néhány információ elveszik, ugyanis nem tudjuk, hogy a köztes csúcsba beérkező és onnan kimenő folyamok közt hogyan oszlanak el a hallgatók.

Ezen az ábrán látszik szemléletesen az emelt érettségi hatása. (2019.)

**Modellezés**

Miután megtörténtek a kellő adatfeldolgozási lépések, nekiláthattunk a modellezésnek is.

Az algoritmusok használatához még egy kicsit módosítanunk kellett a fő adathalmazunkat. Először kiszűrtük a ’Szak’ attribútum kivételével az összes szöveges értékű attribútumot, valamint eltávolítottuk a redundáns oszlopokat is (pl. az ’Összpont’ és ’Össz teljesítmény’ oszlopokat, hisz ezek a többi oszlopbeli értékből számolhatóak). Ezután két modelltípust hoztunk létre: az egyikben együtt vizsgáltunk minden hallgatót, a másikban különválasztottuk a vegyészmérnököket és a biomérnököket. Sajnos környezetmérnök hallgatóból nem volt annyi, hogy elég jól használható, általános modellt tudjunk rájuk építeni. Továbbá az együttes modellben, mivel nagyon kevés 1-es osztályzat született, SMOTE túlmintavételezési eljárással megpróbáltuk dúsítani az adathalmazt. A szétbontott modellben ezt nem tudtuk megtenni, mert ott csak 1 vagy 2 ilyen címkéjű rekord volt, így azt nem tudtuk volna lekezelni SMOTE-val. Emiatt számolnunk kell azzal, hogy a szétbontott adatmodellen használni kívánt algoritmusok emiatt kissé nehézkesen fognak működni.

Első lépésként kNN algoritmust használtunk. Mivel szerepeltek erősen korrelált adatok is az adathalmazban, így az euklideszi távolság mellett Mahalanobis távolsággal is próbáltunk modellt építeni. A megfelelő szomszédszám meghatározásához 10-szeres keresztvalidációt alkalmaztunk mindkét esetben, ahol a szomszédszámot 1-től 30-ig iterálva néztük.

Ezután bináris döntési algoritmusokat próbáltunk ki *One VS Rest* alapon. Mindkét adatmodellen azonos algoritmusokat vizsgáltunk, majd összevetettük őket, hogy leellenőrizzük, nyertünk-e azzal, hogy szétbontottuk az adathalmazt.

* Először SVM algoritmust alkalmaztunk. Révén, hogy az adathalmazban kevés diák kapott elégtelen osztályzatot, paraméternek beállítottuk, hogy kiegyenlítetlen osztályozási feladatról lesz szó. Ezután meg kellett határozni a regularizációs tag értékét. 1-től 5-ig 0.1-es beosztással választottuk ki iterálva a megfelelő értéket, és megnéztük 10-szeres keresztvalidációval, hogy az így kapott modell hogy teljesít (5 felett már nem volt nagy változás pontosság tekintetében). Ahol a legmagasabb volt az átlagos pontosság, azt választottuk.
* Ezt követően logisztikus regresszióval próbálkoztunk. Itt is rögzítettük, hogy kiegyensúlyozatlan osztályproblémával állunk szemben, majd megpróbáltuk helyesen beállítani a paramétereket (hibatag, számítási függvény), azonban azt tapasztaltuk, hogy az alapértelmezett értékekkel lesz a legjobb a modell.

Majd kipróbáltuk a lineáris regressziót is, ami bár folytonos célváltozó prediktálására alkalmas a leginkább, úgy alakítottuk át a végeredményt, hogy a becsült értéket kerekítettük, majd megnéztük, ez hogyan illeszkedik a tényleges teszthalmazbeli értékekre. Az átalakítás előtt még megnéztük az és *RMSE* értékeket.

Végezetül még egy Naive Bayes algoritmust is futtattunk.

A végső, letisztított adathalmaz, amin dolgoztunk, alapból 12 attribútumos rekordokból áll, így azt is megnéztük, hogy főkomponens analízist használva kevesebb, származtatott attribútummal hogyan változik a teljesítmény. Először néhány, intuíciónak megfelelő számú főkomponenssel vizsgáltuk, majd végül végigiteráltunk a lehetséges főkomponens értékeken és megnéztük, hogy adott számú főkomponenssel az eddig használt algoritmusok egy része hogyan teljesít, és ezeket egy táblázatba gyűjtöttük.

Ahogy majd a kiértékelésnél látni fogjuk, ezek változó pontosságú, de még nem túl jó végeredményt adtak nekünk. Intuitívan is érződik, hogy csak egyetem előtti és kora egyetemi hallgatói adatokból nagyon pontos végeredményt nem várhatunk, hiszen az első félév során minden hallgató másképp tud megbirkózni az egyetem által rá nehezedő nehézségekkel, és ez a teljesítményeken is tetten érhető: Valaki nagyon könnyen belerázódik az új, egyetemi környezetbe, így képes ugyanolyan jól teljesíteni, mint középiskolában, valaki nehezebben abszolválja, és emiatt gyengébben teljesít, illetve akadhat olyan is, aki a választott szakon sokkal jobban tud teljesíteni, mint ahogy eddig tette. Azonban helytelen lenne azt feltételezni, hogy semmilyen mértékű következtetést nem tudunk levonni ezekből az adatokból, hiszen általában drasztikus változások nincsenek a hallgatók teljesítményében. Így következő lépésnek így módosítottuk a célváltozóértékeket: eddig öt lehetséges értékünk volt, az ötféle osztályzatnak megfelelően. Ezekből *3 új osztályt* hoztunk létre, egyet a 4-es 5-ös hallgatóknak (ők a jó tanulók), egyet a 2-es, 3-as tanulóknak (ők a gyengén teljesítő tanulók), és az utolsót az 1-es érdemjegyű hallgatóknak (akik nem tudták teljesíteni a tárgyat). Így megváltoztatva a célváltozóértékeket, az előzőekhez hasonlóan, kétféle modelltípuson dolgoztunk, szakonként szétválasztva és egyben. Majd ezekre is kipróbáltuk a korábban használt algoritmusokat, és megnéztük, hogy a pontosság mennyivel változott.

**Kiértékelés**

Összességében négy adatmodell-en vizsgáltuk a kNN, lineáris regressziós, Naive Bayes illetve *One VS Rest* elvű bináris osztályozó algoritmusaink teljesítőképességét attól függően, hogy volt-e szakonkénti bontás, illetve, hogy hány lehetséges értéke volt a célváltozónak. Nem meglepő módon, típustól függően meglehetősen eltérő eredményeket kaptunk.

Együttes, 5 célváltozóértékkel:

Text

Description automatically generated- Összességében nem olyan rossz, de még korántsem optimálisak ezek a megoldások. Érdemes azonban hozzátenni, hogy a véletlen osztályozó ezen a modellen 0.2-es várható pontosságú (hisz 5 lehetséges célérték van), szóval ahhoz képest nem rosszak. Főkomponens analízissel azt kaptuk, hogy 4-5 főkomponens használatával az SVM pontossága 0.44-0.45-re növelhető, de a többi esetben nem tudunk javítani.

Szakonként, 5 célváltozóértékkel:  
*Vegyészmérnökökre:* *Biomérnökökre:*

Text

Description automatically generatedText

Description automatically generated

- Vegyészmérnököknél látható, hogy a kNN osztályozó és a lineáris regresszió teljesítménye romlott, viszont a bináris osztályozók és a Naive Bayes teljesítménye javult. Feltehető, hogy ha rendelkezésünkre állt volna minden jegyhez megfelelő mennyiségű rekord, akkor ez a javulás drasztikus jobb lenne. Főkomponensek használatával azonban nem tudunk javítani.  
- Biomérnököknél viszont minden osztályozó teljesítménye romlott, azonban 3-4 főkomponens használatával a logisztikus regresszió pontossága 0.4-ig növelhető.

Text

Description automatically generatedEgyüttes, 3 célváltozóértékkel:  
  
- Ahogy sejtettük, így jelentősen nőtt a minden osztályozó pontossága, a leglátványosabban talán a bináris osztályozóké, a Bayes-é és a lineáris regreszióé. 4-5 főkomponens használatával az SVM osztályozó pontossága még növelhető 0.7-re, a többi algoritmus azonban nem, vagy csak nagyon kicsiny mértékben javítható.

Text

Description automatically generatedSzakonként, 3 célváltozóértékkel:  
*Vegyészmérnökökre:* *Biomérnökökre:*  
Text

Description automatically generated

- Mindkét esetben az tapasztalható, hogy a kNN osztályozók pontossága kis mértékben csökkent, azonban a többi osztályozó pontossága nőtt. Így megéri predikció előtt megvizsgálni, hogy az adott hallgatót jelölő rekord ’Szak’ attribútuma az milyen értékű. 3-4 főkomponens használatával ráadásul a logisztikus- illetve lineáris regresszió pontossága a vegyészeknél 0.86-ra, míg a biomérnököknél 0.8-ra növelhető.

**Az eredmények vizuális kiértékelése**

A többféle algoritmus többféle adatmodellen való teljesítményének összevetésére rengeteg lehetőségünk van. Az algoritmusok pontosságai modellenként így alakultak:

Chart, bar chart

Description automatically generated

Itt is nagyon jól látszik, hogy a 3 célváltozóértékes modelleknél minden algoritmus pontossága nagyban javult.   
Ugyanakkor ezeket az adatmodelleket akár valamilyen szinten sorba is rendezhetjük a szétbontottságuk és az egyszerűségük szerint, ami tükrözi azt is, hogy mi mely adatmodellekre gondoltunk először, és ezekből kiindulva melyekre később, így mint egy folyamatábraként nézhetjük, hogy hogy változtak a teljesítménymutatók:

Chart, line chart

Description automatically generated

Chart, bar chart

Description automatically generatedÉrdemes lehet esetleg külön vizsgálni, hogy az egyes modelleken mely algoritmusok teljesítettek jól, és melyek kevésbé:

Chart, bar chart

Description automatically generated- Látható, hogy bár tényleg az egyik legegyszerűbb osztályozó algoritmus, mindegyik modellen a többihez képest nagyon jól teljesít az euklideszi távolságfogalommal ellátott kNN osztályzó. Szintén egész jó eredményt értünk el a Naive Bayes-i osztályzóval, illetve a Mahalanobis-os kNN algoritmussal.

A picture containing background pattern

Description automatically generatedChart, bar chart

Description automatically generated- Az SVM és a regressziós algoritmusok a többihez képest gyengébben teljesítettek, viszont a kiértékelésnél tárgyaltaknak megfelelően a pontosságuk javítható főkomponensek használatával, sőt, a potenciálisan legjobb pontosságú osztályozómodellt a legutolsó, 3 csoportos ’Szak’ szerint szétbontott modellben a 3 főkomponensű *One VS Rest* elvű logisztikus regresszióval tudjuk elérni. Ugyanakkor ebben az esetben kicsit nehéz értelmezni a pontos modell együtthatóit és paramétereit egyrészt a főkomponensek miatt, másrészt a *One Vs Rest* elvű osztályozás jellege miatt. (az még egyszerűen kinyerhető, hogy melyik főkomponensben mely attribútumok játszanak fontos szerepet: az elsőben az ’Érettségi pont’, a másodikban a ’Matekeredmény’, a harmadikban a ’Matekeredmény’ és ’Többletpont’ attribútumok.)

Chart

Description automatically generatedVégül nézzük, hogy a pontosan kiértékelt algoritmusok és adatmodellek közül melyik adja a legjobb modellt:  
Nem meglepő módon, a 3 csoportos modelleken tesztelt algoritmusok jobban teljesítenek, mint ahol az 5 jegy volt a célváltozóértékünk. Itt is látszik, hogy a legjobb 3 algoritmus-adatmodell pár mind kNN alapú, utána vegyesen következnek a többiek.

**Következtetések levonása, használhatóság**

Kora egyetemi és felvételi adatok alapján nehezen tudunk közel pontos predikciót adni a hallgatók egyetemi matematika teljesítményére vonatkozólag. Ha a pontos jegyre vagyunk kíváncsiak, ezen módszerekkel kicsivel több mint 50%-os eséllyel tudjuk eltalálni, így javításnak itt még bőven van helye. Viszont, ha nem akarjuk ennyire precízen meghatározni, és csak az érdekel minket, hogy jól vagy rosszul fog teljesíteni az adott hallgató, akkor 80-85%-os pontossággal tudunk dolgozni, ami már nem egy rossz eredmény, így ezen algoritmusok segítségével be tudjuk határolni, hogy mely hallgatók fogják valószínűleg jól venni az akadályokat, és kik azok, akik esetlegesen felzárkóztatásra, segítségre szorulnának. Utóbbi hallgatók számára így az egyetem szervezhet közös korrepetálásokat, feladatmegoldó szemináriumokat, amelyekkel azok a készségek és ismeretek fejleszthetők, amelyekre az egyetemi matematikai elmélet és gyakorlat elsajátítása során szükség van. Ugyanakkor jogosan merülhet fel a kérdés, hogy ez tényleg teljesen etikus-e, hogy akár már azelőtt beskatulyázunk egy hallgatót a ’rosszul teljesítő’ kategóriába, mielőtt bármit is tett volna az egyetemen. Mielőtt ezt az etikai dilemmát nem tisztázzuk, nem igazán használhatjuk ezen algoritmusokat tiszta lelkiismerettel, azonban ha sikerül ezzel kapcsolatban közös megegyezést alkotni a leendő hallgatók és az egyetemi vezetés között, amely lehetővé tenné a modelljeink használatát, akkor hosszútávon valószínűleg csökkenteni tudjuk a lemorzsolódást, és jó előrejelzést tudnánk adni a hallgatók teljesítményére, melynek felhasználásával hasznos változásokat lehetne hozni a tantervben.

**Hivatkozások**

1. Bodnár, G., Berezvai, Sz., Verasztó, Zs., Szilágyi, B.: Hatékony új módszerek alkalmazása a BME reguláris matematika oktatásában, 2016  
   <http://opuseteducatio.hu/index.php/opusHU/article/view/101/113>
2. Novel prediction test for freshmen at BME, Faculty of Chemical Technology and Biotechnology Brigitta Szilágyi, Gábor Hornyánszky, Szabolcs Berezvai, 2019