**Önálló kutatási feladat mérföldkő beszámoló**

**Cikkek tanulságai:**

A miénkhez hasonló tematikájú, de nem pontosan olyan célú projektekről vannak cikkek birtokunkban. BME-n folytatott projektek beszámolóit is megkaptuk, szintúgy Szilágyi Brigitta Tanárnő vezetésével készültek. Ezen tanulmányok fő feladata a gyakori tesztelés, illetve a nemrégiben bevezetett új oktatási módszerek eredményességének vizsgálata, illetve a kognitív és matematikai tudást mérő felmérés alapján ezen képességek közötti összefüggések feltárása. A konklúzió általánosságban a bevezetett módszerek (edubase, online konzultációk, segédanyagok stb.) és a gyakori, illetve gyakorlatias tesztelés jótékony hatását bizonyítja (ezt leginkább kérdőívek és végső eredmények alapján származtatják), emellett kimutatják a korrelációt a matematikai és a kognitív képességek között. Mi is ugyanezen felmérő adatait vizsgáljuk, azonban célunk az utóélet és további eredményesség vizsgálata lesz, a megfelelő összefüggések kimutatása mellett. A cikkekből merítettünk mérési és probléma megközelítési ötleteket, így hasznosnak mondhatjuk a tanulmányozásukat.

**Adattisztítás:**

Az adatokat Szilágyi Brigitta Tanárnő bocsátotta a rendelkezésünkre, amik szerencsére már egészen tiszta állapotban voltak, azonban végeznünk kellett még pár lépést, hogy valóban konzisztens, hiánytalan, használható adatokat kapjunk. Voltak adataink 2019, 2020 és 2021-ről, pontosabban az azévben kezdett diákok kognitív teszteredményeiről. Először el kellett távolítanunk néhány redundáns, illetve irreleváns oszlopot (pl. edubase jelszó, felhasználónév). Ezután volt egy kis inkonzisztencia abban, hogy egyes hallgatók hogyan nevezték meg a szakot, amin tanulnak (kis- és nagybetű, különbség a ’mérnök – mérnöki’ között), úgyhogy ezt pár Python kóddal kiküszöböltük és egységes nevére hoztuk őket (Vegyészmérnöki, Biomérnöki, Környezetmérnöki). Voltak olyan rekordok, ahol nem tudtuk egyértelműen beazonosítani, hogy mi az adott hallgató szakja (mert pl. VBK-t írt oda), így azt átírtuk UNKNOWN-ra, amit esetleg majd a szak alapján csoportosított elemzéseknél elhagyunk (ebből szerencsére évenként csak 2-3 db volt). Kicsit kellett még tisztítani az irányítószámokon, illetve a tankör megnevezésén, de mivel ebből kevés volt, ezt manuális is meg tudtuk tenni.

Az adatok nagyon alaposak voltak, minden hallgatónál külön fel volt jegyezve, hogy mely kérdésre milyen választ adott, és hogy az helyes volt-e. A nagy átfogó elemzéseknél ez nem annyira fontos, ezért azokat különválasztottuk, és áthelyeztük egy másik Excel táblába, amikkel majd a részletesebb elemzéseknél fogunk esetleg foglalkozni. Így ami maradt a 3 évi táblában: a hallgatók neve; Neptun kódja; irányítószáma; emeltezett-e matekból vagy sem; természettudományi tagozatos volt-e vagy sem, illetve a teszteredmények lebontva matekos és nyelvi részre egzakt pontszámmal és százalékos teljesítménnyel. A matek- és kognitíveredmények a 2019-es és 2020-as táblában explicit ki voltak írva, a 2021 táblában csupán az összteljesítmény volt, így ott még cellafüggvényekkel meg kellett határozni a kérdésekre adott válaszokból. A Covid-19 által sújtott években ezen felül még rendelkezésünkre állt egy olyan kategorikus változó is, ami azt fejezte ki, hogy 12. osztályban milyen formában folyt a matematika oktatás. Ezután csináltunk néhány táblát, amiben a különböző évek eredményeit összeillesztettük, nyilván hozzáadva minden rekordhoz egy új oszlopot, amiben az évszám áll, így majd évenkénti bontásban is meg tudjuk vizsgálni az adatokat.

Idő közben kaptunk még 0. Zh eredményeket és felvételi pontszámokat is egyes hallgatókhoz, akik az elmúlt két évben iratkoztak be, így azokat merge-eltük a megfelelő évi táblákkal, és kiszűrtük a többletoszlopokat (mert pl. minden sorban kétszer lenne a hallgató neve).

**Első Adatelemzési lépések:**

Először Tableau segítségével készíttettünk néhány kétdimenziós ábrát arra vonatkozóan, hogy milyen összefüggés van a matematikai és a kognitív pontszámok között. Azt tapasztaltuk, hogy egy minimális pozitív korreláció van a két eredmény között, ám nem túl szignifikáns. Ezután megnéztük évre és szakra lebontva is. Ekkor azt tapasztaltuk, hogy a 2020-as matematika eredmények ugyan jobbak a 2019-eshez képest minimumpontszám alapján, ám összességben nem sokkal. Ennek az lehet az oka, hogy nem nagyon lehet kipuskázni a kérdéseket, mert nem lexikális tudást mérnek. Ugyanakkor a 2021-es eredmények rosszabbak lettek, mint a 2020-as és 2019-es, és ez valószínűleg a digitális időszak következménye. A szakonkénti bontásnál pedig azt tapasztaltuk, hogy a vegyészmérnökök átlagosan jobban teljesítenek mindkét területen, mint a bio- vagy környezetmérnökök. Azt is észrevettük, hogy a pontszámok közti szórás 2021-ben a legnagyobb. Ezután azt vizsgáltuk, hogy mekkora kihatása van a teszteredményekre annak, hogy az adott hallgató emelt matematika érettségit tett-e vagy sem, illetve, hogy természettudományos tagozaton volt-e. Nem meglepő módon az emelt vizsgát tevők sokkal jobban teljesítettek a matekos részen, a kognitív részen viszont sok különbség nem volt. Hasonló volt a helyzet a természettudományos tagozati vizsgálatkor. Az sem lepett meg minket, hogy a felvételi pontszám, a 0. ZH eredménye és a többi teszteredmény között valamekkora pozitív korreláció azért van (a számolt R^2 statisztika 0.2-0.4 közé esett).

Ezekből az szűrhető le, hogy nem véletlenszerűek az eredmények, vannak összefüggések: érdemes külön vizsgálni az emeltes és nem emeltes csoportot, illetve szakonkénti bontásban is.

**Modellezés:**

A célunk eredetileg a hallgatók utóéletének prediktálása (lemorzsolódik-e vagy sem, milyen átlaga lesz stb) volt az eddigi adatok felhasználásával. Sajnos a végső átlagok még hiányosak, mert Szilágyi Brigitta Tanárnő még nem kapta meg mindet, így ott nem tudtunk teljeskörű modellezésbe belefogni, azonban egyes tárgyakból már voltak végső eredményeink, és azokra már tudtunk modellt építeni. Ezekből egy példa az Általános Kémia., a többi tárgynál is hasonlóan jártunk/járunk el.

Általános Kémiánál mivel nem volt túl sok adatunk (kb 300 darab), így a véletlen erdővel nem próbálkoztunk, mert az kicsi adathalmazon nem teljesítene túl jól. Először a jegyre akartunk egy modellt építeni, ehhez döntési fát használtunk. Először felbontottuk az adathalmazt tanító- és teszthalmazra, 70-30 százalék arányban.   
Ezután tanítóhalmazra futtattunk egy döntési fát, aminél nagyon meghatározó volt, hogy egyrészt 180 pont fölött vagy alatt írta az illető az érettségi vizsgát, illetve, hogy melyik tankörhöz tartozott. Ezután megnéztük úgy, hogy a tanköröket kiszűrtük. Ekkor helyette az volt domináló (nem annyira meglepően), hogy az illető milyen szakhoz tartozott (pl. ha valaki 180 pont feletti érettségit írt, az 75% eséllyel kapott ötöst, de aki 180 alatt és még biomérnöknek is tanul, az 65% eséllyel kapott 1-est).  
Az adatfelderítésnek megfelelően megnéztük, hogy szakonként hogy változik a fa. Vegyészmérnököknél a ponthatár 185-re emelkedett, és a második legfontosabb tényező a kognitív teszt matekeredménye lett, biomérnököknél pedig 175-re csökkent a pont. Ugyanígy néztük meg emeltes és nem emeltes csoportra is.  
Ezután visszamértük az eredményeket a teszthalmazon. Mint az várható volt, ha nem bontjuk fel és vizsgáljuk pl. szakokra, akkor az eredmény rosszabb lett (50%-os pontosság), mint ha külön bontjuk (62%-os pontosság). Ugyanakkor ezek még kezdetleges modellek voltak, törekszünk arra, hogy keresztvalidációval meghatározzuk az optimális famagasságot és a használandó attribútumokat.

Még futtattunk egy lineáris regressziót is a jegyek meghatározására. Nem meglepő módon itt is az érettségi pont dominált, illetve ami még jelentős volt, az a többletpontok mennyisége, illetve a kognitív teszt kitöltésére szánt idő. Az R^2 statisztika értéke 0.42 volt, és a teszthalmazon visszamérve is elég nagy volt a szórás, így van még mit csiszolni a modellen (forward selection-nel próbálunk majd javítani). Megvizsgáltuk még a reziduális tagokat is, amik egyenletesen helyezkedtek el, szóval ebből több új információt nem tudtunk kinyerni.

Amit még tervezünk csinálni, hogy többféle adattáblára futtatjuk ezeket az algoritmusokat, ugyanis például elég sok mindenkiről van felvételi információnk, viszont közülük nem mindenki csinált kognitív tesztet. Így, ha például csak a felvételi + egyéb néhány plusz információt vesszük figyelembe, és a kognitív eredményeket nem vesszük hozzájuk, ugyan elvesztettünk néhány attribútumot, de nőtt az adatpontjaink száma, így lehet jobb modellt tudunk találni. Ebből kifolyólag az is előfordulhat, hogy ezen modellel esetleg pár hallgatónál, akiknél az eredmény nem annyira egyértelmű, egyértelműbbé tudjuk tenni a kognitív teszteredmények segítségével.