

# Jövőbe látó lélektan – Gépi tanulás a pszichológiai kutatómódszertanban

Damsa Andrei<sup>1,2\*</sup> , Püski Marcell<sup>3</sup> és Tulics Miklós Gábor<sup>4</sup>

<sup>1</sup> Neuron Solutions, Magyarország

<sup>2</sup> Pécsi Tudományegyetem Bölcsészettudományi Kar Pszichológia Intézet, Alkalmazott Pszichológia Doktori Program, Pécs, Magyarország

<sup>3</sup> Computacenter, Magyarország

<sup>4</sup> Universal Diagnostics, Spanyolország

## EREDETI KÖZLEMÉNY

Beérkezett: 2021. július 20. • Elfogadva: 2023. szeptember 10.

Megjelent az interneten: 2024. március 7.

© 2023 A szerző(k)



**Háttér és célkitűzés:** A pszichológiai kutatómódszertan eljárásait (főképp a  $p$  értékre építkező bizonyításokat) számos kritika érte az utóbbi évtizedek során. A kutatói elfogultság és a módszertanok (például az adatgyűjtés, az adatszelekció vagy a statisztikai próbák) könnyű manipulálhatósága teret adott a félrevezető és nehezen reprodukálható kutatásoknak. A gépi tanulás elterjedése megfigyelhető a pszichológia területén is, új eszköztárat biztosítva a kutatók számára. Az eljárás áthelyezi a hangsúlyt a statisztikai bizonyításról az előrejelzésre, valamint az ehhez kapcsolódó validációs folyamatokra, ezáltal lecsökkentve a kutatói szubjektivitás hatását. Jelen tanulmány célja gyakorlati példákon keresztül betekintést nyújtani a gépi tanulás módszertanába, fókuszálva a pszichológiai alkalmazhatóságára. **Módszer:** A vizsgálati szakasz első részében két, a gépi tanulás használatára irányuló tanulmány kerül bemutatásra a humán döntéshozatali mechanizmusok, valamint a pandémiás helyzet okozta mentális hatások területére vonatkozóan. A vizsgálati szakasz második részében egy klasszifikációs feladat (filmpreferencia és nemi identitás kapcsolata) keretén belül kerül összehasonlításra egy nem parametrikus statisztikai módszer és két, gépi tanuláson alapuló eljárás. **Eredmények:** A kapott eredmények bemutatják a gépi tanulás által nyújtott előnyöket (validációs eljárások és többletinformáció kinyerése), párhuzamot vonva a nem parametrikus eljárással. **Következtetések:** A tanulmány népszerűsíteni és alátámasztani hivatott a gépi tanulás alkalmazhatóságát a kutatói szektorban tevékenykedő pszichológusok számára. A bemutatott kutatás reprodukálhatóságának érdekében az adatok és programozási kódsorok szabadon felhasználhatók a tanulmányban megadott elérhetőségeken keresztül.

\* Levelező szerző. E-mail: damsa.andrei@neuronsolutions.hu

## KULCSSZAVAK

gépi tanulás,  $p$  érték,  $p$ -hacking, pszichológiai kutatómódszertan, klasszifikáció

## EMPIRIKUS MEGKÖZELÍTÉS A PSZICHOLÓGIÁBAN – A $P$ ÉRTÉK SEBEZHETŐSÉGE

A pszichológia célja feltérképezni és megérteni a viselkedés mögött meghúzódó érzelmi, valamint kognitív folyamatokat. Az emberi természet tanulmányozása építkézhet szubjektív tapasztalatok értelmezésére, induktív következtetésekre vagy empirikus eszközök alkalmazására. Utóbbit támogatva született meg az a matematikai eljárásokat alkalmazó kutatómódszertan, amely az adatok és az azokra irányuló statisztikai próbák segítségével tár fel, támaszt alá vagy cáfol meg jelenségeket. Annak ellenére, hogy ez a módszertani megközelítés erőteljes löketet adott a pszichológia tudományos elismertségének, a publikált pszichológiai vizsgálatok során megmutatkozó statisztikailag szignifikáns eredmények sok esetben félrevezető és nehezen reprodukálhatók lehetnek (Open Science Collaboration, 2015). A pszichológiai kutatómódszertan tekintetében számos kritika érte (és éri napjainkban is) a  $p$  értékre épülő bizonyítási eljárásokat (Wicherts, Veldkamp, Augusteijn, Bakker, Van Aert és Van Assen, 2016). A hipotézisvizsgálat esetében a  $p$  érték annak a valószínűséget jelenti, hogy a tesztstatisztika értéke legalább akkora, vagy nagyobb, mint a mintából kapott érték (azzal a feltevessel, hogy igaz a nullhipotézis). A standard eljárások szerint a nullhipotézist (ami azt mondja ki, hogy az összehasonlított értékek között nincs különbség, változás vagy eltérés) elvethetjük, ha a  $p$  érték 0,05-dal egyenlő, vagy az alatti. Ebből kifolyólag a hipotézisvizsgálatok során az egyik cél lehet az, hogy a  $p$  értéket minél alacsonyabban tartsuk (fontos kiemelni, hogy ez az érték nem egyenlő a hipotézis validitásának vagy erejének mértékével). Simmons, Nelson és Simonsohn (2011) vizsgálata bemutatja, hogy erősebb kutatói hajlam irányul olyan (esetenként hamis) bizonyítékok felhasználására, amelyek alátámasztják a preferált hipotézist, mint az olyanokra, amelyek megcáfolják azt. Ezt erősíti meg az a tendencia is, mely szerint a tudományos publikálási folyamatban a lektorálást és szakmai minősítést végző szakemberek, illetve maguk az olvasók is azokat a tudományos műveket részesítik előnyben, amelyek hipotéziskongruensek, tehát megfelelnek az előzetesen felállított kutatói elvárásoknak (Ioannidis, 2012).

A kutatói elfogultság megjelenésére számos lehetőség adódik a kutatási kérdés felmerülésétől a végleges publikációig, amelyek a vizsgálatot végző szakember preferenciája és elvárásai alapján torzíthatják az eredményeket, ezáltal pedig a kutatásból levont következtetéseket. Ezt a jelenséget nevezzük a  $p$  érték manipulálásának („ $p$ -hacking”-nek), mely során a vizsgálat különböző lépéseiben az elvárt eredmények által vezérelve történnek kisebb-nagyobb torzítások (Simmons, Nelson és Simonsohn, 2011). Strube 2006-os tanulmánya kimutatta, hogy a kutatásuk keretében megkérdezett, szakmai publikációkat végző pszichológusok 56%-a nyilatkozta azt, hogy az adatgyűjtés során figyelembe vette a  $p$  értékeket, ami befolyásolta a minta kialakítását. Továbbá a kutatásban részt vevők 46%-a mondta azt, hogy csak olyan vizsgálatokat publikált, ahol beigazolódott az előzetesen felállított hipotézis, és 38% nyilatkozta azt, hogy az adatok (például a változók) kiválasztásánál figyelembe vette az elvárt eredményeket. Ahogy Yarkoni és Westfall (2017) is megemlíti, érdemes elgondolkodni azon, hogy ezek a számok önbevallásos alapú adatgyűjtés során alakultak ki, így a valódi értékek feltehetően ennél magasabbak lehetnek.



A kutatási eljárás kialakításakor, valamint a vizsgálatok elvégzése során számos olyan pont van, amely esetében megjelenhetnek a végső eredményt torzító hatások (Head, Holman, Lanfear, Kahn és Jennions, 2015). Az ilyen jellegű problémák elkerülése végett ajánlott olyan módszertani megoldásokat alkalmazni, amelyek lehetőséget adnak az adatgyűjtési folyamat ellenőrzésére és a statisztikai modellek objektív visszamérésére, ezáltal rávilágítva az eljárások esetleges pontatlanságára. A gépi tanulás módszertanára alapozó megközelítések több szempontból is hasznosnak bizonyulhatnak a pszichológiai kutatómódszertan szempontjából, egyrészt predikciós erejük, másrészt a részletes és szabályozott adat-előkészítési és modellezési folyamataik miatt (Yarkoni és Westfall, 2017).

## GÉPI TANULÁS

A pszichológia tekintetében a gépi tanulást főképp a Cambridge Analytica köré felépülő botrány hozta be a köztudatba, amely rávilágított arra, hogy az online közösségi hálókön elérhető publikus információk alapján magas fokú pontossággal lehet prediktálni olyan személyes jellemzőket, mint például a vallás vagy a politikai nézet. A Cambridge Analytica munkáját megalapozó, Kosinski, Stillwell és Graepel (2013) által végzett kutatás során kirajzolódott, hogy a pszichológiában addig még nem, vagy csak kevésbé használt gépi tanulási eljárások nagymértékben segíthetik a tudományterület fejlődését. Amint azt Orrù, Monaro, Conversano, Gemignani és Sartori (2020) megfogalmazták, a gépi tanuláson alapuló pszichológiai eljárások új szemléletet hozhatnak be a kutatók számára, áthelyezve a hangsúlyt a statisztikai következtetésről a gépi tanulási modellekre építkező előrejelzésre.

Amint azt a cikk elején bemutattuk, a pszichológia tudományának célja az emberi viselkedés megértése, mely magában foglalja a viselkedés értelmezését és a viselkedéshez kapcsolódó predikciók megfogalmazását (Yarkoni és Westfall, 2017). A napjainkban megfigyelhető kutatási tendenciák esetében azonban kevés hangsúly tevődik a viselkedés előrejelzésére, a vizsgálatok nem fordítanak elég figyelmet az elméletek alkotása mellett azok gyakorlati felhasználására (Yarkoni és Westfall, 2017). Ennek oka a múltban feltehetően az volt, hogy nem állt a kutatók rendelkezésére megfelelő eszköz, amely hatékonynak bizonyult volna a komplex emberi viselkedés sikeres modellezésére. Napjainkban viszont, a gépi tanulás és egyéb technológiák fejlődésével, valamint a hatalmas mennyiségű, könnyen elérhető adatok felhasználásával ez már egy valóságos cél (Jacobucci és Grimm, 2020). A gépi tanulás predikcióalapú megközelítése hatékony gyakorlati kiegészítése lehet a – többnyire hipotézistesztelésen alapuló – kutatásoknak, segítve azok értelmezhetőségét, relevanciáját, érvényességét és gyakorlati felhasználását.

A gépi tanulás a mesterséges intelligencia azon ága, amely a tanulni képes rendszerekkel foglalkozik (Larose és Larose, 2014). Ezek a rendszerek felismerik az adatokban meghúzódó mintázatokat, melyek alapján következtetéseket tesznek meglévő vagy új adatok alapján. A gépi tanulás eljárásait három fő kategóriába sorolhatjuk annak függvényében, hogy milyen céllal alkalmazzuk őket.

1. **Nem felügyelt tanulás:** Az adatokban rejlő mintázat és a feltételezett eredmény ismeretlen, az algoritmus feladata megtalálni az esetleges információt, az adatpontok közötti hasonlóságok alapján (például klaszterezés).
2. **Felügyelt tanulás:** A tanítási szakaszban ismert a célváltozó, azonban az azt leíró mintázat ismeretlen, az adatokból kiszűrve ezt az algoritmus találja meg (például osztályozás).



3. **Megerősített tanulás:** Az algoritmus feladata megtalálni a legjobb stratégiát egy adott cél elérése érdekében (például egy önvezető autó esetében).

A feltáró jelleggel bíró vizsgálatok mellett a pszichológiai kutatómódszertan esetében a felügyelt módszerek kategóriájába tartozó eljárások lehetnek kiemelten hasznosak. Az ide sorolandó megoldások két fő területre irányulnak:

1. **Osztályozási (vagy klasszifikációs) feladatok:** A célváltozó nominális, az algoritmus feladata, hogy az adatok alapján eldöntse az adott megfigyelésről, hogy milyen kategóriába (osztályba) tartozik (például nő vagy férfi).
2. **Regressziós feladatok:** A célváltozó skála (intervallum) típusú, az algoritmus az adatok alapján becsüli meg a célváltozó értékét (például egy adott teszt pontszámát).

Amennyiben a vizsgált célváltozó ordinális, a problémát kezelhetjük akár osztályozási, akár regressziós feladatként is.

A statisztikai következtetésre épülő modellekkel szemben a gépi tanulás nem az előre megfogalmazott feltételezésekre, hanem az adatokban lévő mintázatokra épít. Ebből kifolyólag az algoritmusnak nincsenek előzetes elvárásai a végeredményt illetően, az elemzés során a hangsúly sokkal inkább az adatok előkészítésén van. Ennek ellenére az elméleti pszichológiai tudás (*domain knowledge*) a gépi tanulás módszertanával végzett kutatásban is elengedhetetlen, például a megfelelő adatforrások megválasztása és az eredmények értelmezése során (Elhai és Montag, 2020). Mivel a hangsúly a tanulási fázisban rendelkezésre álló adatokon van, ezért fontos a felhasznált változók mennyisége, valamint a minta elemszáma. A modellek ellenőrzésekor is lényegi különbségek mutatkoznak meg a hipotézisvizsgálat során használt eljárásokkal szemben – a gépi tanulás esetében a pontosságot az határozza meg, hogy a modell milyen eredményességgel tud előre jelezni (kategorizálni vagy becsülni) olyan adatok mentén, amelyeket a tanulási szakaszban nem ismert.

## GÉPI TANULÁS A PSZICHOLÓGIAI KUTATÁSMÓDSZERTANBAN

A következőkben olyan kutatásokat mutatunk be, amelyek a klasszikus kutatómódszertani megközelítések mellett gépi tanulás segítségével vizsgáltak különböző pszichológiai jelenségeket. Az első kutatás (Rosenfeld, Zuckerman, Azaria és Kraus, 2012) döntésméleti kérdéskörben alkalmazta a gépi tanulás megközelítését, a második (Flesia és mtsai, 2020) pedig a pandémiás helyzettel kapcsolatos mentális hatások vizsgálatának esetében használta a módszertant.

### Emberi döntéshozatal és gépi tanulás

Rosenfeld és munkatársai 2012-ben végzett, a területen még újnak számító kutatása jól demonstrálja a gépi tanulás és a pszichológiai modellek együttes használatának előnyeit. A kutatók tanulmányukban az emberi döntéshozatal vizsgálatán keresztül hasonlították össze a témakörhöz kapcsolódó pszichológiai megközelítések és gépi tanuláson alapuló modellek predikciós pontosságát. A kutatás különböző helyzetekben vizsgálta a résztvevők döntéshozatali viselkedését, majd döntésspszichológiai modellek (Korlátozott racionalitás elmélete, Fókuszpont elmélet), gépi tanuláson alapuló modellek, illetve a kettő együttes használatán alapuló hibrid modellek segítségével próbálta meg lemodellezni azokat. A modellek hatékonyságát a kutatók két döntési helyzetben alkalmazták, az első egy egyszerű optimalizációs feladat volt (egy vásárlás esetében



kellett meghatározni, hogy mikor hagyják abba a legmegfelelőbb bolt keresését a kísérleti személyek), a második pedig egy komplexebb, tárgyalási helyzet volt (a résztvevők egy számítógéppel játszottak, és pénzüsszegek közül kellett kiválasztaniuk ugyanazt az értéket, mint a program). A modellek azt hívatottak prediktálni, hogy ezekben a helyzetekben hogyan döntenek a kísérleti személyek.

A vizsgálat egyik központi kérdésköre az volt, hogy hogyan lehet kombinálni a társadalomtudósok által javasolt kognitív döntéshozatali megközelítéseket a gépi tanulás módszertanával. A kapott eredmények azt mutatják, hogy egyszerű döntési helyzetekben, mind a pszichológiai megközelítések, mind a gépi tanuláson alapuló modellek megbízható előrejelzést eredményeznek, közel azonos pontossággal. Ezekben a szituációkban – amennyiben elegendő adat áll rendelkezésre – a gépi tanuláson alapuló modellek a kutatók által elfogadhatónak tekintett pontossággal képesek prediktálni a résztvevők viselkedését, és a hibrid megközelítések (a kognitív pszichológiai modellek és a gépi tanulási eljárások ötvözei) nem javítanak szignifikáns módon az előrejelzés pontosságán. Ettől függetlenül megmutatkozott egy fontos különbség a gépi tanulási modellek és a hibrid megközelítés között – utóbbi esetben kevesebb adat felhasználása is elegendő volt az azonos eredmény eléréséhez. Összetett döntési helyzetekben sem a pszichológiai megközelítések, sem a gépi tanuláson alapuló modellek nem szolgáltak megbízható predikciós értékekkel. Ezekben a helyzetekben a cikk által bemutatott újszerű hibrid megközelítés – a gépi tanulási algoritmusokat alkalmazva (melyek ebben az esetben az osztályozásra irányultak) a megoldás alapjaként, de a kognitív modellek jellemzőinek hozzáadásával – jelentős javulásokat eredményezett mindkét alapvető (pszichológiai modellek és gépi tanulás) megközelítéssel szemben.

## A pandémiás helyzet mentális hatásainak vizsgálata gépi tanúlással

Flesia és munkatársai (2020) kutatása is szemléltető gyakorlati példája annak, hogy hogyan lehet a gépi tanulást pszichológiai vizsgálatokban alkalmazni. A módszertant a hagyományos statisztikai modellek eredményeinek kiegészítése és értelmezése mellett a kutatók gyakorlati prediktív eszközként is használták. A kutatás célja a SARS-Cov-2 (Covid-19) vírus és annak elterjedése által kialakult pandémiás helyzet mentális egészségre gyakorolt negatív hatásainak feltárására irányult, valamint kitért azoknak a személyeknek az azonosítására is, akik magas rizikófaktorúnak számítanak a vírus okozta stresszterhelés szempontjából. A kutatásban különböző szociodemográfiai jellemzők (például kor, nem, a háztartásban élő családtagok és/vagy más személyek száma, havi háztartási jövedelem stb.) és pszichológiai változók (megküzdés, önkontroll, kontrollhely, személyiség stb.) hatását vizsgálták a járványhelyzethez kapcsolódó stressz érzékelésére. Az összhatások feltérképezése mellett a kutatás azt is célul tűzte ki, hogy prediktív modellek segítségével azonosítani tudja azokat a magas rizikófaktorral bíró csoportokat, akikből a vírus emelt szintű stresszt válthat ki (a magas rizikófaktorú csoportok detektálása különösen fontos a különböző prevenció és stresszcsökkentő eljárások mielőbbi és hatékony alkalmazása szempontjából). A „hagyományos” elemzés során a szerzők többváltozós regresszióanalízist futtattak az érzékelt stressz skála pontszám és a független változók közötti kapcsolat vizsgálatára annak érdekében, hogy meghatározzák azokat a változókat, melyek szignifikánsan hozzájárultak az érzékelt stressz szintjéhez. A kutatók a magas rizikójú személyek azonosítása érdekében gépi tanuláson alapuló klasszifikációs modelleket használtak (logisztikus regresszió, szupport vektor gép, Naiv Bayes, véletlen erdők), melyek két csoport mentén (alacsony és magas stresszszint) tanultak rá az adatokban lévő mintázatokra (az elemszám – 2053 fő – 80%-át használták a



modellek tanítására, a maradék 20%-át pedig a tesztelésre). Az alkalmazott modellek a teszt-mintán megközelítőleg 0,76-os pontossággal teljesítettek a magas érzékelt stresszszinttel rendelkező személyek azonosításában, fókuszálva a helyes besorolás és az elsőfajú hiba kategóriájába eső személyekre (a pozitív esetekben különösen fontos lehet a megelőző beavatkozások kidolgozása). A többváltozós regresszióanalízis által meghatározott változók és a gépi tanuláson alapuló módszer közös használatával a magas stressz rizikóval jellemezhető személyek azonosítása mellett lehetőség nyílt a hatékonyabb stresszcsoökkentő eljárások kidolgozására is, melyek – a kapott eredmények alapján – főképp az önszabályozási készségek fejlesztésére koncentrálnak.

A bemutatott két kutatás jól demonstrálja, hogy a gépi tanulás alkalmazása nem feltétlenül jár minden esetben bonyolult módszertannal, valamint viszonylag egyszerű felépítésű kutatásoknál is hasznosan lehet alkalmazni az eredmények szélesebb körű felhasználásának érdekében.

## GYAKORLATI PÉLDA A GÉPI TANULÁS ALKALMAZÁSÁRA

A gépi tanulás folyamatának szemléltetése érdekében egy egyszerű vizsgálaton keresztül mutatjuk be a módszertant, az elemzés lépéseit tartalmazó Jupyter Notebook (Python, v. 3.8.8) parancssor, valamint az alapul szolgáló szabadon elérhető adatbázis megtalálható a megadott GitHub címen.<sup>1</sup>

### A vizsgálat elméleti hátterének rövid bemutatása

A szemléltető vizsgálat a Kaggle.com<sup>2</sup> weboldalon található, fiatalok és fiatal felnőttek bevonásával készült adatokat elemzi, fókuszálva a filmpreferencia nemi eltéréseire. Az adatfelvétel 2013-ban zajlott, a kutatásban 1010 fő vett részt, az adatbázis tisztítása után a végleges elemszám 980 fő lett (kikerültek a hiányzó értékeket tartalmazó esetek).

A filmpreferencia területét a pszichológia számos kérdéskör mentén vizsgálja, ilyen például a nemi sztereotípiák vagy a médiahasználati szokások megértése (Wühr, Lange és Schwarz, 2017). Az eddigi kutatások eredményeiből azt a következtetést lehet levonni, hogy megjelennek kimutatható különbségek a férfiak és a nők filmpreferenciáját illetően – a nők a férfiakhoz képest jobban preferálják a pozitív érzelmi hangulatú filmeket, így például jobban kedvelik a romantikus és a vígjáték kategóriákat (Greenwood és Lippman, 2009). A férfiak a nőkhöz viszonyítva jobban kedvelik a magasabb aktivációs (*arousal*) szintet kiváltó filmeket, továbbá a férfiak a nőkhöz képest jobban élvezik az olyan filmek nézését, amelyekben megjelenik az agresszió – ennek fényében az általuk kedvelt kategóriák például az akció- és a horrorfilmek (Wühr, Lange és Schwarz, 2017). Ezeknek a nemi különbségeknek a meghatározása fontos az olyan elméleti kérdések megértéséhez, amelyek kiterjednek a férfi-női nemi különbségekre, a sztereotípiák vizsgálatára és az eltérő médiafogyasztási szokások azonosítására (Redfern, 2012).

### A vizsgálat előkészítése

Az adatfelvételi szakaszban a kitöltők 1-től 5-ig tartó Likert-skálán jelölték a preferencia mértékét adott témakörökön belül (1 – egyáltalán nem preferált, 5 – teljes mértékben preferált).

<sup>1</sup>[https://github.com/andreadamsa/jovobe\\_lato\\_lelektan.git](https://github.com/andreadamsa/jovobe_lato_lelektan.git)

<sup>2</sup><https://www.kaggle.com/miroslavsabo/young-people-survey>





Az elemzés során az adatbázisból kiemelésre kerültek a filmpreferenciával kapcsolatos változók (11 filmkategoría, a kategóriák listáját az [1. melléklet](#) tartalmazza) és a nemi eloszlást mutató oszlop (férfi = 40%, nő = 60%). A kiválasztott változók mentén azt vizsgáltuk, hogy milyen eltérést mutat a filmpreferencia a nem függvényében, illetve prediktálható-e egy adott kitöltő neme az alapján, hogy milyen stílusú filmeket preferál.

A vizsgálat során két megközelítést alkalmaztunk, az első esetben nem parametrikus statisztikai próba segítségével térképeztük fel a preferencia eloszlását a filmkategoróriák esetében a nem mint csoportképző változó tekintetében. A második esetben gépi tanulás alkalmazásával készítettünk két modellt, amelyek célja megismerni és predikciós helyzetben alkalmazni a nem és a filmpreferencia közötti kapcsolatot.

### Filmpreferencia vizsgálata nem parametrikus statisztikai próba alkalmazásával

A kutatás első szakaszában Mann-Whitney U próba alkalmazásával vizsgáltuk meg a különböző filmstílus-kategóriákhoz kapcsolódó preferenciák különbségeit a nem mint csoportosító változó függvényében. A Mann-Whitney U próbát rendszerint a Student-féle t próba alternatívájaként használják azokban az esetekben, amelyekben a t próba előfeltételei (például az adatok normális eloszlása) nem teljesülnek (Fay és Proschan, 2010). A normalitás vizsgálatát a Shapiro-Wilk-teszt (Razali és Wah, 2011) segítségével végeztük el, mely esetében egyik változó mentén sem teljesültek a normális eloszlás feltételei (a kapott eredményeket a [2. melléklet](#) tartalmazza). Ebből kifolyólag ebben a kutatási szakaszban a további vizsgálatok során a Mann-Whitney-próbát alkalmaztuk.

Két csoport összehasonlítása során a nullhipotézis azt feltételezi, hogy a vizsgált változó mentén az eloszlások azonos mintázatot mutatnak – amennyiben ez nem bizonyítható, meg kell vizsgálni az alternatív hipotézis által kínált magyarázatot. Jelen kutatás esetében a nullhipotézis szerint az adatok mentén nem mutatható ki eltérés a filmpreferencia területén a nemi csoportosítás függvényében.

A Mann-Whitney U próba értelmezése során két fő mutatót használtunk fel, az egyik az U, a másik a  $p$  érték. Kisebb U érték esetén várhatóan nagyobb lesz az eltérés a csoportok által adott válaszok között, a  $p$  érték pedig a statisztikai szignifikancia meghatározásához szükséges ( $\alpha = 0,05$ ). A próba alkalmazásánál megmutatkozott, hogy egy filmkategoría kivételével minden esetben szignifikáns különbség rajzolódott ki a nemi eloszlás függvényében (az eredményeket a [3. melléklet](#) foglalja össze). A legmarkánsabb különbség (a legalacsonyabb U érték mentén) a romantikus ( $U = 57688$ ;  $p < 0,001$ ; férfi átlag = 2,84; női átlag = 3,92) és a háborús filmek ( $U = 59858$ ;  $p < 0,001$ ; férfi átlag = 3,83; női átlag = 2,68) esetében jelent meg. Ezekkel ellentétben a komédia kategóriába tartozó filmek esetében nem mutatkozott meg statisztikailag szignifikáns eltérés a két csoport átlagpontszámai között ( $U = 113590$ ;  $p = 0,290$ ; férfi átlag = 4,48; női átlag = 4,51).

A Mann-Whitney U próba segítségével helyzetképet kaptunk a filmpreferencia és a nemi hovatartozás közötti kapcsolatról, viszont – a szignifikancia alkalmazásán kívül – nehezen tudjuk megítélni, hogy a kinyert információk a gyakorlatban milyen mértékben alkalmazhatók, így a kutatás validitása is megkérdőjelezhető. Annak érdekében, hogy a kapott információk gyakorlati hasznát feltérképezzük, gépi tanulás segítségével vizsgáljuk meg a filmpreferencia nemi vonatkozásait.

### Filmpreferencia vizsgálata gépi tanulás algoritmusok alkalmazásával

A Mann-Whitney U próba segítségével megmutatkozott, hogy a vizsgált filmkategoróriák többségében statisztikailag szignifikáns eltérés található a nemi eloszlás mentén a preferencia



kérdéskörében. A továbbiakban részletesebben megvizsgáljuk ezt a kérdéskört, és gépi tanulás alkalmazásával határozzuk meg a nem mint csoportosító tényező (célváltozó), valamint a különböző filmkategóriákhoz köthető preferencia (bemeneti változók) kapcsolatát. A vizsgálat során két felügyelt klasszifikációs eljárást alkalmaztunk, az első a logisztikus regresszió, a második a K legközelebbi szomszéd módszertanára épül.<sup>3</sup>

A kutatásban szereplő célváltozót (amely esetében a mintázatok feltérképezi az algoritmus) a nem képezte (0 = nő, 1 = férfi, az osztályok közötti eloszlás kiegyensúlyozottnak mondható, férfi = 40%, nő = 60%, összesen 980 fő), a bemeneti változókat pedig a filmkategóriák alkották (az előzőleg leválogatott 11 változó). A gépi tanulás alkalmazásának első lépései közé tartozik az adatok megfelelő előkészítése, amely során olyan támpontokat kell figyelembe venni, mint például a hiányzó értékek kezelése, a kiugró esetek kiszűrése, valamint a nem megfelelően bevitt adatok újrakódolása/értelmezése. A gépi tanulás egyik, széles körben alkalmazott keretrendszere a CRISP-ML<sup>4</sup> (*Cross-industry Standard Process for Machine Learning*), amely lépésről lépésre leköveti az adatelemzési folyamat menetét, kitérve a különböző alfeladatokra. A CRISP-ML által nyújtott szemléletmód vonalvezetőként funkcionálhat a kutatók számára, segítve a reprodukálható, validált és átlátható elemzési folyamatot.

A jelen kutatás során használt adatbázis letisztított és jól strukturált formában érhető el, így a vizsgálat első szakaszában csak a hiányzó értékeket kezeltük – mivel relatív kevés ilyen eset volt (33 darab, az adatok 3%-a), ezeket kivettük az adattáblából (ennek értelmében az összes olyan sor törlésre került, amelyben volt hiányzó érték). A tanítási szakasz futtatása előtt – a változók hatásainak megfelelő értelmezésére vonatkozóan – megvizsgáltuk a bemeneti változók lehetséges összefüggéseit. Ezt a multikollinearitás fogalomkörében tisztáztuk (Maddala és Lahiri, 2009), mely a változók között fennálló, a célváltozó szempontjából félrevezető értelmezéssel bíró kapcsolatokat tárja fel. A varianciainflációs tényező (Variance Inflation Factor – VIF) számítása során megmutatkozott, hogy a vizsgált bemeneti változók között nem áll fenn multikollinearitás (a változókat 5-ös VIF érték esetében érdemes tovább vizsgálni, jelen kutatásban 2,08 volt a legmagasabb VIF érték a fantázia és mese filmkategóriában).

A következő lépés a klasszifikációs algoritmusok betanítása volt, amely során a már említett logisztikus regressziót, illetve a K legközelebbi szomszéd algoritmust alkalmaztuk. A vágási pontot (*threshold*) – mely alapján dönt az algoritmus az osztály besorolásáról – mindkét eljárás esetében az alapértelmezett 0,5-ös értéken hagytuk. Szemléltetésképpen a 4. melléklet mutatja be a vágópont változtatásának hatását a gyakoribb metrikákra a logisztikus regresszió esetében.

## Logisztikus regresszió alkalmazása

A logisztikus regresszió hasonló módon működik, mint a lineáris regresszió, azzal a lényegi különbséggel, hogy a célváltozó nem intervallum, hanem nominális típusú. Az algoritmus a bemeneti változók mentén úgy változtatja a saját paramétereit, hogy a lehető legnagyobb pontossággal becsülje meg az adatsorokhoz tartozó bináris osztályt (jelen esetben ez a férfi vagy nő kategória). A gépi tanulás módszertanában fontos szerep jut a validációnak, avagy az algoritmus által prediktált eredmények visszaellenőrzésének. Ebből kifolyólag a betanítási és visszaellenőrzési szakasz két külön adattáblán történik azért, hogy az algoritmust egy „számára új” helyzetben

<sup>3</sup><https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article>

<sup>4</sup><https://ml-ops.org/content/crisp-ml>





próbáljuk ki. A rendelkezésre álló adattáblát a legtöbb esetben 20%–80% vagy 30%–70% felosztásban használják, a tanító adatbázis javára (Bronshtein, 2017). Jelen esetben a 30%-70%-os felosztást alkalmaztuk, mely során a tanító adatbázisba 686, a tesztelő adatbázisba pedig 294 kitöltés került. A tanító adatbázisban összesen 277 férfitől és 409 nőtől, a tesztelő adatbázisban pedig összesen 119 férfitől és 175 nőtől származó minta szerepel.

A tanító adatbázisban szereplő bemeneti változók és célváltozó mentén elvégeztük a betanítást, majd felhasználva a tesztadatáblát, megvizsgáltuk a modell validitását. A tesztadatbázis esetében az algoritmus a tanítási szakaszban meghatározott paraméterezés mentén a bemeneti változók felhasználásával becsüli meg az osztályhoz való tartozást (férfi vagy nő a kitöltő) anélkül, hogy ismerné az adott sorokra vonatkozó célváltozó értékét (a nemet). A validitás vizsgálata a tévesztési mátrixon keresztül két fő mutató mentén történt – az első a pontosság (*Accuracy*), a második a Matthews-féle korrelációs együttható (*Matthew's corcoef* – *MCC*) volt, mindkettő az algoritmus által becsült eredményeket hasonlítja össze a tesztadatbázisban szereplő osztályozási értékekkel. Az MCC figyelembe veszi a valódi és hamis pozitív, valamint negatív eredményeket (*True és False positives/negatives*), ezért kiegyensúlyozott mérőszámnak tekinthető, amely akkor is használható, ha az osztályok nagyon különböző méretűek (a pontosság ebben az esetben félrevezető lehet). A pontosságon és az MCC metrikán kívül érdemes még megemlíteni a precizitást (*precision*) és a fedést (*recall*) mutatóit is. A fedés metrikája megadja a pozitív osztály esetében helyesen prediktált értékek arányát. A precizitás arra a kérdésre válaszol, hogy ha a modell igennel prediktál, akkor az hányszor helyes. Az orvostudományi kutatásokkal kapcsolatosan gyakori a specificitás (*specificity*) és az érzékenység (*sensitivity*) mutatóinak kiszámítása az orvosi tesztek értékelésénél. A fogalmak nagyon hasonlóak az előzőekhez, de nem szabad szem elől téveszteni a különbségeket. A specificitás az igaz negatív arány, vagyis a valódi negatív eredmények aránya minden olyan esethez képest, amelyeknek negatívnak kellett volna lenniük, az érzékenység pedig a helyesen prediktált pozitívumok arányát méri. Látható, hogy a fedés és az érzékenység megegyezik, azonban a precizitás és a specificitás különbözik.

A modell validációja során elsőként a tévesztési mátrixot vizsgáljuk meg, amely az 1. táblázatban szereplő eloszlást mutatja.

A táblázatban szereplő bal felső és jobb alsó mezők (*True positive és True negative*) mutatják azokat az eseteket, amelyekben az algoritmus által becsült érték egybeesik a valós értékkel. A bal alsó és jobb felső mezők (*False positive és False negative*) szerepeltetik a hibásan becsült kategóriákat, ide tartoznak azok az esetek, amelyekben az algoritmus nő esetén férfi, férfi esetén pedig női osztályt rendelt az adott kitöltéshez. A tévesztési mátrixból kiszámolva a pontosság értéke 0,79 lett, a Matthews-féle korrelációs együttható pedig 0,58-as értéket vett fel. Összehasonlítva hasonló jellegű kutatásokkal (például előzőleg prezentáltakkal, Flesia és mtsai, 2020; Rosenfeld és mtsai, 2012), ezek az eredmények elfogadhatónak számítanak. Jelen vizsgálatban a logisztikus regresszióra épülő algoritmus a validációs szakaszban 294-ből 234 esetben becsülte

**1. táblázat.** Tévesztési mátrix a nemre vonatkozóan logisztikus regresszió alkalmazása során (becsült érték – az algoritmus által meghatározott érték, valós érték – az adattáblában szereplő eredeti érték)

<i>n</i> = 294	Becsült érték: nő	Becsült érték: férfi
Valós érték: nő	141	34
Valós érték: férfi	26	93

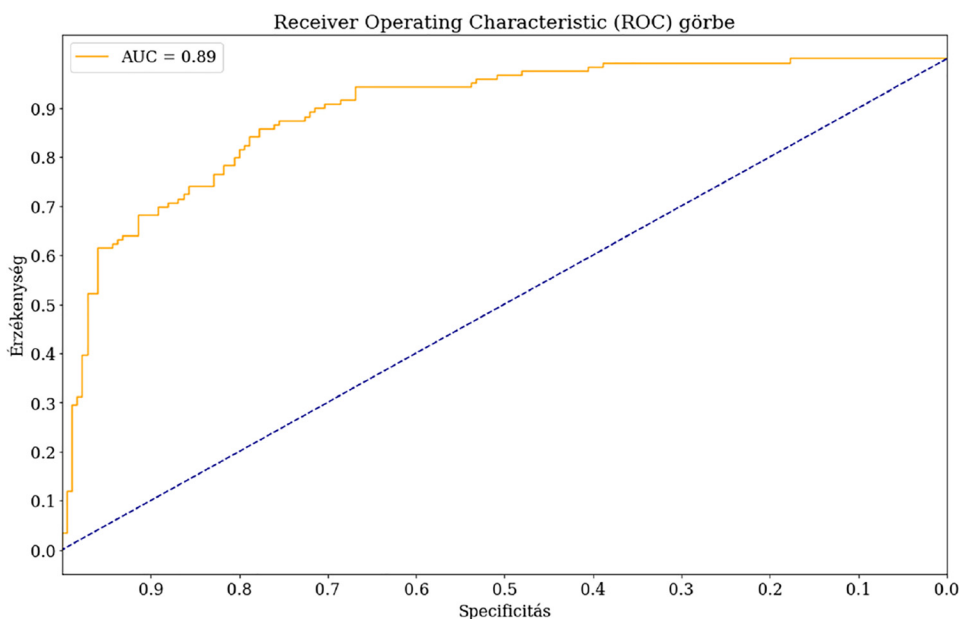


meg helyesen a kitöltő nemét az alapján, hogy a különböző filmkategóriák esetében milyen preferenciaértéket mutatott.

A pontosság vizsgálata során a görbe alatti terület (*Area Under the Curve* – AUC) és a vevő működési karakterisztika (*Receiver Operating Characteristics* – ROC) mutató is meghatározásra került, amely az osztályozási problémák teljesítményét jeleníti meg különböző vágópontok mellett. A ROC egy valószínűségi görbe, az AUC pedig a szeparálhatóság mértékét mutatja meg, felvázolja, hogy a modell milyen mértékben képes megkülönböztetni a célváltozóban szereplő osztályokat. Minél magasabb az AUC, annál hatékonyabban képes a modell a 0 osztályokat 0-nak, az 1-es osztályokat pedig 1-nek prediktálni. Ennek értelmében egy ideális AUC-ROC görbe az ábra bal felső sarka felé tart. Jelen vizsgálat esetében az AUC értéke 0,89, amely magas osztályozási képességre vall (a görbét az 1. ábra mutatja be).

A felvázolt metrikák vizsgálata mellett a modell működésébe való betekintés is értékes lehet a kutatási kérdéskör feltérképezése és értelmezése során. Ennek egyik módja az algoritmus által felhasznált bemeneti változók fontosságának vizsgálata (*feature importance*), avagy annak értelmezése, hogy az algoritmus milyen súllyal használta fel az adott bemeneti változót a becslés folyamatában. Az elemzésünk során az algoritmus a 2. táblázatban szereplő bemeneti változók esetében ítélte meg a legnagyobb súlyozást (a táblázatban az első öt változót tüntettük fel, a teljes lista az 5. mellékletben található).

Az algoritmus által használt változók fontosságának (súlyozásának) vizsgálatából jól látszik, hogy a női-férfi preferencia különbség olyan filmek esetében mutatkozik meg, amelyek például a romantikus, az akció vagy a háborús kategóriákba esnek. Ezek az eredmények egybevágnak a Mann–Whitney U próba alkalmazása során kapott értékekkel, valamint a filmpreferencia nemi



1. ábra. ROC görbe és AUC érték a logisztikus regresszió modell eredményei alapján



**2. táblázat.** Az algoritmus által használt legnagyobb súllyal rendelkező első öt bemeneti változó (a pozitív előjelű változók a férfi címke felé tolják a predikciót, a negatív előjellel rendelkezők pedig a női címke irányába)

Sorszám	Változó neve (filmkategória)	Súlyozási érték	<i>p</i> érték
1.	Romantic	−0,966	0,000
2.	Action	0,584	0,000
3.	War	0,514	0,000
4.	Western	0,477	0,000
5.	Fantasy/Fairy tales	−0,433	0,001

különbségeivel foglalkozó elméleti bevezetőben felvázolt kutatásokkal (Greenwood és Lippman, 2009; Wühr, Lange és Schwarz, 2017). A vizsgálat során nem mutatkozott meg statisztikailag szignifikáns eredmény a horror, thriller, animációs és dokumentumfilmek esetében, ezek elhagyása növelheti a prediktív modell pontosságát.

## K legközelebbi szomszéd algoritmus alkalmazása

A logisztikus regresszió bemutatása mellett a gépi tanulás egy másik, gyakran megjelenő módszere is alkalmazásra kerül. A K legközelebbi szomszéd az adatpontok között fennálló hasonlóságokra fókuszál, ezeket tanulja meg a célváltozó vetületében. Az algoritmus egy új adatpont klasszifikációja során megvizsgálja az adatpont *k* számú legközelebbi szomszédját, majd a szomszédok osztályai alapján tesz becslést az adatpontra. Például, ha az adatpont három legközelebbi szomszédja közül kettő az „A”, egy pedig a „B” osztályba tartozik, az algoritmus az „A” osztályba sorolja be a pontot a predikció során. Fontos szem előtt tartani, hogy a vizsgált szomszédok száma páratlan kell legyen, valamint ne egyezzen meg a lehetséges osztályok számával.

Az algoritmus betanítása és tesztelése során az előző részben szereplő adathalmazt használtuk, az előzőekben felvázolt azonos eloszlási arányokkal. Az eljárás során az alapértelmezett paraméter-beállítások voltak felhasználva, melyeket a modellt tartalmazó scikit-learn<sup>5</sup> (sklearn) Python modul definiálja. Ezek alapján a keresett legközelebbi szomszédok száma 5, a távolságok kiszámítására pedig az euklideszi távolság került meghatározásra. A gépi tanulási eljárások esetében az adatok megfelelő előkészítése mellett fontos szerep jut a modellekben megtalálható paraméterek beállításának is. Jelen tanulmánynak nem célja a paraméterek vizsgálatának bemutatása, azonban fontos megemlíteni, hogy számos, specifikus technika (Chauhan, 2020) foglalkozik a kérdéskörrel.

Annak ellenére, hogy az algoritmus működési mechanizmusa eltérő a logisztikus regresszió megközelítéstől, a kapott eredmények a vizsgált adatbázis esetében azonosak. Mind a használt metrikák (a pontosság értéke 0,79, a Matthews-féle korrelációs együttható pedig 0,58), mind a tévesztési mátrix során identikus eloszlást és számokat találunk az algoritmus futtatását követően (ebből fakadóan ezek részletes bemutatása az előző részben megtörtént). Ez a jelenség jelen esetben az adatbázis sajátosságaiból fakadóan adódhat (ugyanazokra a mintázatokra tanult rá a két algoritmus), viszont eltérő helyzetekben a két eljárás különböző eredményt hozhat.

<sup>5</sup><https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>



Akárcsak a klasszikusnak számító statisztikai próbák esetében, a gépi tanulás algoritmusok kiválasztása során is megfontolandó a vizsgált probléma felvetése, a használt adatbázisok tulajdonságai, valamint a számítási kapacitás rendelkezésre állása. A logisztikus regresszió futási ideje a vizsgált adatbázis esetében 34,3 milliszekundum, a K legközelebbi szomszéd során 14,6 milliszekundum. Jelen esetben ezek az időarányok nem számottevők, viszont egy komplexebb és nagyobb adatbázis esetében (akár mélytanulási eljárásokat alkalmazva) a tanítási idő és kapacitás fontos tényezővé válhat.

A felvázolt metrikákon kívül a logisztikus regresszió és a K legközelebbi szomszéd tesztelési adatokon adott predikciói is összehasonlításra kerültek, ezek alapján a vizsgált mintán nem volt eltérés a két modell osztályozási eredményeiben. A gépi tanulási modellek kiválasztásánál azonban további szempont lehet az értelmezhetőség, például a logisztikus regresszió esetében viszonylag egyszerűen kimutatható a bemeneti változók súlyozása, mely a betanított modell működési logikájába ad betekintést. A K legközelebbi szomszéd alkalmazása során ezek a tényezők ilyen módon nem nyerhetők ki, ugyanis az algoritmus a pontok közötti távolságok alapján dönt – egy megfelelően normalizált vagy skálázott adathalmazban a mérvadó a pontok egymáshoz viszonyított helyzete lesz, így a változók egyformán fontos szerepet játszanak a lokalizálásban. A gépi tanulási eljárások implementálása során figyelembe kell venni az algoritmusok működési elvéből származó sajátosságokat és ezek fényében alkalmazni őket. Ezen okok miatt a K legközelebbi szomszéd algoritmust olyan alkalmazásokhoz használhatjuk, amelyek nagy pontosságot igényelnek, de nem kritikus pont az ember által olvasható modell kialakítása. A gyakorlatban az algoritmust sokszor (más népszerű technikák mellett) alapmodellként használják, amelyhez hozzá lehet hasonlítani a többi gépi tanuló modellt. A modellek által prezentált eredmények érthetősége érdekében léteznek olyan eljárások, amelyek modelldiagnosztikus módon (adott algoritmus működési elvétől függetlenül) magyarázzák a bemeneti változók szerepét a predikció során (ilyen például a SHAP megközelítés, a parciális dependenciák vizsgálata, vagy a LIME; [Nori és mtsai, 2019](#)).

A betanított algoritmusok olyan esetekben is használhatók, amelyek jelenleg nincsenek benne a tanításhoz vagy a validációhoz felhasznált adatbázisokban. Így például, ha egy új kitöltő válaszol a filmpreferenciával kapcsolatos kérdésekre, az algoritmus megbízható pontossággal meg tudja becsülni azt, hogy az illető férfi vagy nő.

## ÖSSZEFOGLALÁS ÉS KITEKINTÉS

A tanulmányban szereplő példák és elemzések két dolgot hivatottak megerősíteni – egyrészt azt, hogy – kiegészítve a klasszikus eljárásokat – a gépi tanulás értékes eszköze lehet a pszichológiai kutatómódszertannak, másrészt azt, hogy alkalmazását ajánlott elterjeszteni a kutatók körében, hiszen a módszertan az egyetemi oktatás keretén belül ugyanúgy elsajátítható, mint a klasszikusnak számító statisztikai próbák. Fontos megemlíteni azonban, hogy a gépi tanulás eszköztárában szereplő eljárások mind hatékonyság, mind alkalmazhatóság szempontjából is eltérőek lehetnek. Jelen tanulmány keretében felhasználásra került a logisztikus regresszió, amely egy egyszerűen értelmezhető algoritmus, ezzel párhuzamosan – a szemléltetés céljából – ugyancsak beemelésre került a K legközelebbi szomszéd eljárás, amely a logisztikus regresszióhoz képest másik logika mentén közelíti meg a problémát, és nem ad olyan könnyen értelmezhető modellt a felhasználó kezébe, mint a logisztikus regresszió.



Az elemzések során alkalmazott algoritmusok és kódsorok rendszerint előkészített szoftver-csomagokban találhatók meg, amelyek szabadon felhasználhatóak – sok esetben anélkül, hogy a kódsorok mélyére tekintve teljes mértékben átlátnánk az algoritmus komplexitását. Hasonlattal élve olyan ez, mint az autóvezetés – más megtanulni szabályosan és biztonságosan vezetni egy járművet, és ugyancsak más autószerelőként vagy autótervezőként dolgozni. Sokan vezetünk úgy autót, hogy nem ismerjük mélységében a jármű működési elvét, mégis hatékonyan tudjuk használni. A gépi tanulás egy rendkívül bonyolult és szerteágazó terület, de alkalmazásához már az a tudás is elegendő lehet, amelyet egy, a szakterületén elmélyedő pszichológus sajátít el.

A filmperefenciával kapcsolatos elemzés során megmutattuk, hogy a gépi tanulás alkalmazása releváns többletinformációval szolgálhat egy adott kérdéskör vizsgálatában. A módszerrel betekintést ad az algoritmus döntési mechanizmusába, például abba, hogy mi alapján kategorizálja be a válaszadókat egy osztály mentén. Ha megismerhető az osztályozási logika, az fontos és értékes információkkal szolgálhat a jelenség értelmezésében – például az algoritmus által használt változók fontosságának (súlyozásának) sorrendje utalhat arra, hogy melyek az osztályozás szempontjából legmarkánsabban megjelenő hatások és tényezők. A filmpreferencia nemi eltéréseinek vizsgálata során például kirajzolódott az, hogy az algoritmus a romantikus, western és háborús filmkategoriókat szerepeltette az első három helyen, a súlyozás előjele (pozitív vagy negatív) pedig megmutatta, hogy a filmkategória preferenciája a női vagy férfi osztály felé tolja el a döntési mechanizmust. Komplex kutatási kérdéskörökben, ahol lényegesen több változóval dolgozik az algoritmus, értékes lehet a releváns változók felismerése és kiszűrése.

A gépi tanulás folyamat szinten is nagymértékben hozzájárulhat a pszichológiai kutatásmódszertanhoz. Az adatok megismerése és megfelelő előkészítése megelőzheti a módszertani hibákat, fókuszálva a lehetséges – akár emberi jellegű – potenciális buktatókra. Az adatbányászatból átvett CRISP-ML<sup>6</sup> keretrendszere strukturált és reprodukálható kutatási folyamatot biztosít, amely csökkentheti a hibák előfordulási arányát. A keretrendszer hat lépést foglal magába, amelyek a pszichológiai kutatásmódszertanra lefordítva a következők.

1. Elméleti háttér megismerése
2. Adatok megismerése és megértése
3. Adatok előkészítése
4. Modellezés
5. Kiértékelés
6. Alkalmazás (gyakorlati kipróbálás)

A gépi tanulás módszertanában markáns szerep jut az adatok előkészítésének, az algoritmusok rendszerint csak olyan adathalmazokkal működnek jól, amelyek a megfelelő formátumban állnak rendelkezésre (Abdallah és mtsai, 2016). Az adatok megtisztítása és kezelése nemcsak a feltáró elemzések folyamatát segíti (zaj kiszűrése, anomáliák detektálása stb.), hanem lehetőséget ad a nem megfelelő adatok (például elírások vagy hiányzó értékek) által okozott torzítások elkerülésére (North, 2015). A keretrendszer által biztosított struktúra, valamint a különböző lépéseknél megjelenő feladatok mederbe terelik a kutatói munkát, csökkentve az emberi hibázás lehetőségét.

<sup>6</sup><https://ml-ops.org/content/crisp-ml>



Amint azt a tanulmány első felében bemutattuk, a  $p$  értéken alapuló statisztikai megközelítés sebezhetővé teszi a kutatási folyamatot. A gépi tanulás során nem egy előre meghatározott kritériumot próbálunk teljesíteni (például a szignifikanciaszint átlépését), hanem szabad teret adunk az algoritmusnak, hogy az adatokban meglévő mintázatok és kapcsolatok mentén találjon értékes információt. A validáció sem egy külső, sok esetben ökölszabály alapján meghatározott érték mentén történik, hanem adatokra építkező visszamérésen (tesztadatbázison). Ennek vitathatatlan előnye, hogy gyakorlati megoldást alkalmaz, ezzel csökkentve az elméleti síkon meghatározott kritériumokból származó torzításokat. Az elemzés továbbá nemcsak egy helyzetképet ad (mint például adott csoportok összehasonlításának esetén), hanem egy előrejelző, becslések generálására alkalmas algoritmust kap a kutató, amelyet felhasználhat prediktív előrejelzések megfogalmazására. A gépi tanulás alkalmazásával a pszichológiai kutatómódszertan egy rendkívül értékes eszközzel bővítheti a tárházát, mely a jelen vizsgálatán kívül lehetőséget ad a jövő modellezésére is.

## IRODALOM

- Abdallah, Z. S., Du, L., & Webb, G. I. (2016). Data Preparation. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 318–327). Totowa: Humana Press.
- Bronstein, A. (2017). Train/Test split and cross validation in Python. Letöltve: 2021. 07. 20-án: <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>.
- Chauhan, N. S. (2020). *Hyperparameter optimization for machine learning models*. Letöltve 2022. 05. 28-án: <https://www.kdnuggets.com/2020/05/hyperparameter-optimization-machine-learning-models.html>.
- Elhai, J. D., & Montag, C. (2020). The compatibility of theoretical frameworks with machine learning analyses in psychological research. *Current Opinion in Psychology*, 36, 83–88.
- Fay, M. P., & Proschan, M. A. (2010). Wilcoxon–mann–whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4, 1–39.
- Flesia, L., Monaro, M., Mazza, C., Fietta, V., Colicino, E., Segatto, B., & Roma, P. (2020). Predicting perceived stress related to the Covid-19 outbreak through stable psychological traits and machine learning models. *Journal of Clinical Medicine*, 9(10), 3350.
- Greenwood, D., & Lippman, J. R. (2009). Gender and media: Content, uses, and impact. In J. C. Chrysler, & D. R. McCreary (Eds.), *Handbook of gender research in psychology* (pp. 643–669). New York: Springer-Verlag.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of P-hacking in science. *PloS Biology*, 13(3), e1002106.
- Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7, 645–654.
- Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, 15(3), 809–816.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behaviour. *PNAS U.S.A.*, 110(15), 5802–5805.
- Larose, D. T., & Larose, Ch. D. (2014). *Discovering knowledge in data: An introduction to Data Mining*. Hoboken: John Wiley & Sons, Inc.
- Maddala, G. S., & Lahiri, K. (2009). *Introduction to econometrics*. Chichester: Wiley.





- Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). *InterpretML: A unified Framework for machine learning interpretability*. Letöltve: 2022. 05. 30-án: <https://arxiv.org/abs/1909.09223>.
- North, M. A. (2015). *Data mining for the masses*. Global Text Project Book.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349.
- Orruù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology*, 10, 2970.
- Razali, N., & Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21–33.
- Redfern, N. (2012). Correspondence analysis of genre preferences in UK film audiences. *Participations: Journal of Audience and Reception Studies*, 9(2), 45–55.
- Rosenfeld, A., Zuckerman, I., Azaria, A., & Kraus, S. (2012). Combining psychological models with machine learning to better predict people’s decisions. *Synthese*, 189(1), 81–93.
- Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.
- Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods*, 38, 24–27.
- Wühr, P., Lange, B. P., & Schwarz, S. (2017). Tears or fears? Comparing gender stereotypes about movie preferences to actual preferences. *Frontiers in Psychology*, 8, 428.
- Wicherts, J. M., Veldkamp, C. L. S., Augusteyn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7, 1832.
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122.

## 1. melléklet: A vizsgálatban szereplő filmkategóriák

Sorszám	Filmkategória
1.	Horror
2.	Thriller
3.	Komédia (Comedy)
4.	Romantikus filmek (Romantic)
5.	Tudományos-fantasztikus (Sci-Fi)
6.	Háborús filmek (War)
7.	Fantáziafilmek (Fantasy and Fairy Tales)
8.	Animációs filmek (Animated)
9.	Dokumentumfilmek (Documentary)
10.	Westernfilmek (Western)
11.	Akciófilmek (Action)



2. melléklet: A Shapiro-Wilk normalitás teszt eredményei

Sorszám	Változó	<i>t</i> érték	<i>p</i> érték
1.	Horror	0,883	<i>p</i> < 0,001
2.	Thriller	0,902	<i>p</i> < 0,001
3.	Comedy	0,672	<i>p</i> < 0,001
4.	Romantic	0,893	<i>p</i> < 0,001
5.	Sci-fi	0,904	<i>p</i> < 0,001
6.	War	0,896	<i>p</i> < 0,001
7.	Fantasy/Fairy tales	0,860	<i>p</i> < 0,001
8.	Animated	0,845	<i>p</i> < 0,001
9.	Documentary	0,885	<i>p</i> < 0,001
10.	Western	0,840	<i>p</i> < 0,001
11.	Action	0,886	<i>p</i> < 0,001

3. melléklet: Mann-Whitney U próba eredményei (Csoportosító változó a nem)

Sorszám	Változó	U érték	<i>p</i> érték
1.	Horror	93937	<i>p</i> < 0,001
2.	Thriller	90484	<i>p</i> < 0,001
3.	Comedy	113591	<i>p</i> = 0,289
4.	Romantic	57688	<i>p</i> < 0,001
5.	Sci-fi	73345	<i>p</i> < 0,001
6.	War	59858	<i>p</i> < 0,001
7.	Fantasy/Fairy tales	77706	<i>p</i> < 0,001
8.	Animated	89807	<i>p</i> < 0,001
9.	Documentary	99225	<i>p</i> < 0,001
10.	Western	68719	<i>p</i> < 0,001
11.	Action	60499	<i>p</i> < 0,001

4. melléklet: Metrikák alakulása a vágópontok változtatásának függvényében a logisztikus regresszió esetében

Vágópont	Pontosság	MCC	Érzékenység	Specifititás
0,100	0,636	0,430	0,975	0,406
0,150	0,704	0,499	0,941	0,543
0,200	0,745	0,557	0,941	0,611
0,250	0,769	0,579	0,916	0,669
0,300	0,789	0,600	0,891	0,720

(continued)



Folytatás

Vágópont	Pontosság	MCC	Érzékenység	Specifitás
0,350	0,793	0,600	0,874	0,737
0,400	0,810	0,623	0,857	0,777
0,450	0,803	0,601	0,815	0,794
0,500	0,796	0,582	0,782	0,806
0,550	0,803	0,589	0,739	0,846
0,600	0,810	0,600	0,706	0,880
0,650	0,806	0,595	0,639	0,920
0,700	0,816	0,626	0,605	0,960
0,750	0,799	0,592	0,563	0,960
0,800	0,789	0,576	0,521	0,971
0,850	0,762	0,521	0,454	0,971
0,900	0,731	0,461	0,370	0,977
0,950	0,680	0,355	0,227	0,989

5. melléklet: Változók súlyozása a logisztikus regresszió által

Sorszám	Változó neve (filmkategória)	Súlyozási érték	<i>p</i> érték
1.	Romantic	−0,966	0,000
2.	Action	0,584	0,000
3.	War	0,514	0,000
4.	Western	0,477	0,000
5.	Fantasy/Fairy tales	−0,433	0,001
6.	Comedy	0,413	0,006
7.	Sci-Fi	0,361	0,00
8.	Animated	−0,157	0,199
9.	Thriller	−0,105	0,352
10.	Documentary	−0,066	0,532
11.	Horror	0,058	0,538

Predictive psychology – applying machine learning in psychological research methodology

Andrei Damsa, Marcell Püski and Miklós Gábrriel Tulics

*Background and aims:* Methods used in psychological research (especially *p*-value based techniques) have been the subject of numerous criticisms in the past decades. Researcher’s subjectivity and easily manipulable methods (regarding data gathering, data selection, or statistical processes) lead to an increased number of misleading and hard-to-reproduce research. Machine learning, as a method, quickly found its



way into the field of psychology, enriching the researchers' toolbox. The methodology shifts the emphasis from statistical explanation to prediction and related validation processes, thereby reducing the impact of researchers' subjectivity bias. The aim of the present study is to provide an insight into the methodology of machine learning through practical examples, focusing on psychological applicability. *Method:* In the first part of the research phase, the use of machine learning is presented through two studies from the areas of human decision-making, and the mental effects of the pandemic situation. In the second part of the research phase, a non-parametric statistical method and two machine learning-based procedure are compared during a classification task (relationship between film preference and gender identity). *Results:* The results demonstrate the benefits of applying machine learning (validation procedures and extraction of additional information) in comparison with the use of non-parametric procedure. *Conclusion:* The study promotes and supports the applicability of machine learning methods in the psychological research sector. To reproduce this research, the data and programming scripts are open-source materials and may be freely used through the links provided in the study.

## KEYWORDS

machine learning, *p*-value, *p*-hacking, psychological research methodology, classification

**Open Access nyilatkozat.** A cikk a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0>) feltételei szerint publikált Open Access közlemény, melynek szellemében a cikk bármilyen médiumban szabadon felhasználható, megosztható és újraközölhető, feltéve, hogy az eredeti szerző és a közlés helye, illetve a CC License linkje és az esetlegesen végrehajtott módosítások feltüntetésre kerülnek. (SID\_1)

