

DEBRECENI EGYETEM
INFORMATIKAI KAR
ADATTUDOMÁNY ÉS VIZUALIZÁCIÓ TANSZÉK



TDK-dolgozat

**Szerencsejáték-függőség prediktálása
klaszterező eljárásokkal**

Témavezető:

Dr. Hajdu András
Egyetemi tanár

Szerző:

Francuz Ádám
Adattudomány MSc
2. évfolyam

Debrecen, 2024

Tartalomjegyzék

1. Szakirodalmi áttekintés	3
2. Paraméterek meghatározása	6
2.1. Paraméterek meghatározása NLP segítségével	6
2.2. Adathalmaz bemutatása	8
2.2.1. Árbevétel	9
2.2.2. TJB	9
2.2.3. Szelvényszám	9
2.2.4. Aktív napok száma	9
2.2.5. Két fogadás között átlagos eltelt idő	9
2.2.6. Kiugró nyereség/veszteség utáni fogadás változás	10
2.2.7. Kommunikáció utáni fogadási szokás változás	10
2.2.8. Heti átlagok megoszlása	10
2.2.9. Fogadási mutatók változása a regisztrációhoz képest	10
2.2.10. Fogadási szünetek	11
2.2.11. Hosszan tartó veszteségsorozatok	11
2.3. Adattisztítás	11
3. Modell létrehozása	12
3.1. Dimenziócsökkentő eljárások	12
3.1.1. Uniform Manifold Approximation and Projection . .	12
3.1.2. t-Distributed Stochastic Neighbor Embedding	15
3.1.3. Saját megközelítés	17
3.2. Végleges klaszterek meghatározása	19
3.3. Ellenőrzés felügyelt gépi tanuló algoritmusokkal	21
4. Elméleti modellezés	23
4.1. Dimenziócsökkentő eljárások modellezése	23
4.1.1. t-SNE modellezése	23
4.1.2. UMAP modellezése	24
4.1.3. Saját megközelítés modellezése	25
4.2. Távolságok modellezése	25
5. Eredmények kiértékelése	27
6. Összefoglalás	28
7. Irodalomjegyzék	30

Bevezetés

A gépi tanuló algoritmusok és mesterséges intelligencia alapú adatmodellek előfordulása az ipari gyakorlatban folyamatosan növekszik. Ezen kutatási területek multidiszciplinárisak, ugyanis az informatikatudomány mellett szükséges a kutatott terület ismerete is. A modellek felhasználásának elterjedése akár évtizedes problémákára adhat megoldást, hiszen a feldolgozható adatok mennyisége sokszorosa a korábban felhasználtaknak. Ezen adatok elemzése lehetővé teszi a rejtett mintázatok és összefüggések felismerését, amelyeket a hagyományos módszerekkel nem lehetett megvalósítani. Az exponenciálisan növekvő technológiai lehetőségek - annak előnyei mellett - teret kínálnak kártékony tevékenységek alkalmazására is, köztük lehetőséget adnak különböző szenvedélybetegségek gyakorlására. Ilyen betegségnek tekinthető a szerencsejáték-függőség, amelyet leginkább a ráfordított idő és elköltött pénz tényezők jellemeznek.

Kutatásom célja egy fogadási mintázatokra épülő modell létrehozása, ami kiterjeszthető és alkalmazható az ipari gyakorlatban. A függőség meghatározására a mai napig nem létezik pontos definíció, így kizárólag felügyelet nélküli tanulási módszereket (Unsupervised Learning) lehet alkalmazni, hiszen olyan modell létrehozása szükséges, ami az adatpontok közötti hasonlóságra összpontosít. Fontosnak tartom, hogy az adattudományi eszközöket és lehetőséget kiterjesszük más tudományterületekre, ugyanis így érhető el a lehető legnagyobb értékteremtés. A függőség prediktálása naprakész terület, a felelős játékszervezési tevékenységek hangsúlya meglehetősen nagy. A függő játékosok deifinálásával nagy áttörés érhető el, a meghatározott ügyfeleknek célzott kommunikációs tevékenységek vagy szolgáltatási megvonások segítségével csökkenteni lehet a függőség mértékét.

A szerencsejáték iparág számos termékből tevődik össze, a legismertebb szolgáltatások a kaszinók, lóversenyek, sportfogadások és lottójátékok. Tudományos munkám során valós, online sportfogadási adatokat használok fel. Az adathalmaz kiválasztása során ügyeltem arra, hogy kellően nagy adathalmazzal tudjak dolgozni, illetve minden külső tényezőtől mentes, például kiemelt sportesemények nélküli időszak adataival dolgozzak, ugyanis ezen események kellően tudják torzítani az adatokat. A sportfogadási adatoknál kiemelt szerepe van az internetes felületnek, mivel ez garantálja az ügyfél-fogadás kapcsolat létrejöttét, tehát a fogadásokhoz köthető ügyféladatokat, ez a kapcsolat elengedhetetlen az aggregált adatok létrehozásában. Fontos kiemelni, hogy csak fogadási adatok állnak rendelkezésre. A függőség prediktálásában fontos adatok lehetnek az adott ügyfél jövedelmi adatai, azonban ezek elérhetetlenek, így célszerű olyan modell létrehozására törekedni, amely elérhető adatokra épül.

Elsőként egy szakirodalmi áttekintés keretein belül vizsgálom a kutatásomhoz köthető meglévő szakirodalmakat. A szerencsejáték-függőséghez köthető irodalmak kifejtik a probléma kialakulását és annak súlyosságát, majd vizsgálom azon tudományos munkákat, amelyek szenvedélybetegség prediktálásához köthető modellek alkalmazásának lehetőségeit mutatják be. Az alapadatokat NLP modell segítségével határozom meg, bemutatom a feldolgozás menetét, a létrehozott promptokat és a végeredményként kapott 18 numerikus paramétert, amelyeknél törekedtem arra, hogy különböző megközelítésekben jellemezze a függőségre való hajlam mértékét. Az adattisztítási folyamat után a modellezés első lépése a dimenziócsökkentő eljárások megvalósítása. A kutatás során a *t-Distributed Stochastic Neighbor Embedding* (t-SNE) és *Uniform Manifold Approximation and Projection* (UMAP) eljárásokkal kétdimenziósra csökkentem az adathalmazt, így már vizualizációk segítségével tudom elemezni az adatpontok közötti kapcsolatokat és összefüggéseket. Ezen két eljárás mellett létrehoztam egy saját megközelítésre alapuló módszert, amely egy numerikus értékkel jellemzi a függőségre való hajlamot.

A két dimenziócsökkentő eljárás után klaszterezem a kétdimenziós adathalmazokat és a fogadási adatokra visszavetítve megállapítom a kezdeti, függő ügyfeleket tartalmazó klasztert. A többi, nem függő klaszterbe tartozó ügyfelekhez létrehozom a klasztertől való euklideszi távolságot, így a saját eljárással kiegészítve ügyfelenként már három megközelítéssel, három távolságértékkel jelölöm a függőségre való hajlamot. A távolságértékeket a *k-Means* algoritmussal klaszterezem, a létrejött klaszterek közül a legkisebb távolságértékekkel rendelkező halmaz tartalmazza a függő ügyfeleket. Ellenőrzésként megvizsgálom az ügyfelek fogadási adatait, ezáltal validálva a létrehozott modell működését. A létrehozott címkék (függő vagy nem függő) felhasználásával bináris osztályozási módszerekkel ellenőrzöm, hogy az újabb adatpontokat milyen arányban lehet helyesen osztályozni, ami igazolja a valós felhasználási lehetőségeket.

A kutatás eredményeképpen, a meglévő adattudományi eszközök felhasználásával és kombinálásával egy olyan módszert hozok létre, amely egy súlyos és növekvő tendenciában lévő multidiszciplináris problémára kínál megoldást. A függőség ilyen szintű és adat alapú prediktálására korábban sem nemzetközi, sem hazai kutatások nem készültek, a valós adatokra való modellépítés pedig garantálja a felhasználási lehetőségeket.

1. Szakirodalmi áttekintés

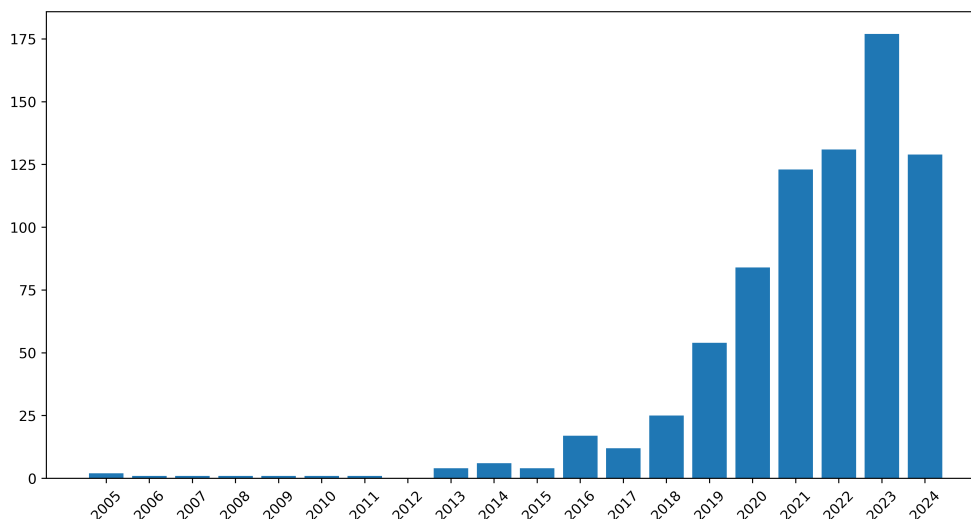
Jelen fejezet célja egy szisztematikus szakirodalmi áttekintés során megismerni a tudományos munkámhoz kapcsolható kutatási előzményeket, a felhasznált modellek alkalmazhatóságának lehetőségeit más iparágakban, illetve azonosítani azon területeket, amelyek tovább fejleszthetők és javíthatóak hasonló tudományos munkákkal. A fejezetben először elemzem a kutatások számának megoszlását különböző paraméterek szerint, majd kiemelem az általam legfontosabbnak vélt kutatásokat a témában. Végül levonom a következtetéseimet a meglévő szakirodalmakkal kapcsolatban.

Az áttekintés során egyenlő figyelmet fordítok a más tudományterületek által készített tudományos munkákra, illetve az informatikatudományi eredményekre, ugyanis a kutatásom is multidiszciplináris problémát dolgoz fel. A szakirodalom megismerésére a *Scopus* adatbázist és az alábbi szűrőket használom:

- TITLE-ABS-KEY ("Machine Learning") AND TITLE-ABS-KEY("Addiction")
- TITLE-ABS-KEY("Data") AND TITLE-ABS-KEY("Addiction") AND TITLE-ABS-KEY("Gambling")

Elsőként a gépi tanuláshoz és függőséghez köthető 774 db cikket dolgoztam fel. Mind a két tudományterület növekvő tendenciákat mutat, amelyek érezhetőek a megjelenő cikkek számában is: ahogyan az 1.1. ábrán is megfigyelhető, a vizsgált szakirodalmak több, mint 80%-a 2020 után jelent meg, az előző évben pedig rekord mennyiségű tudományos munka született.

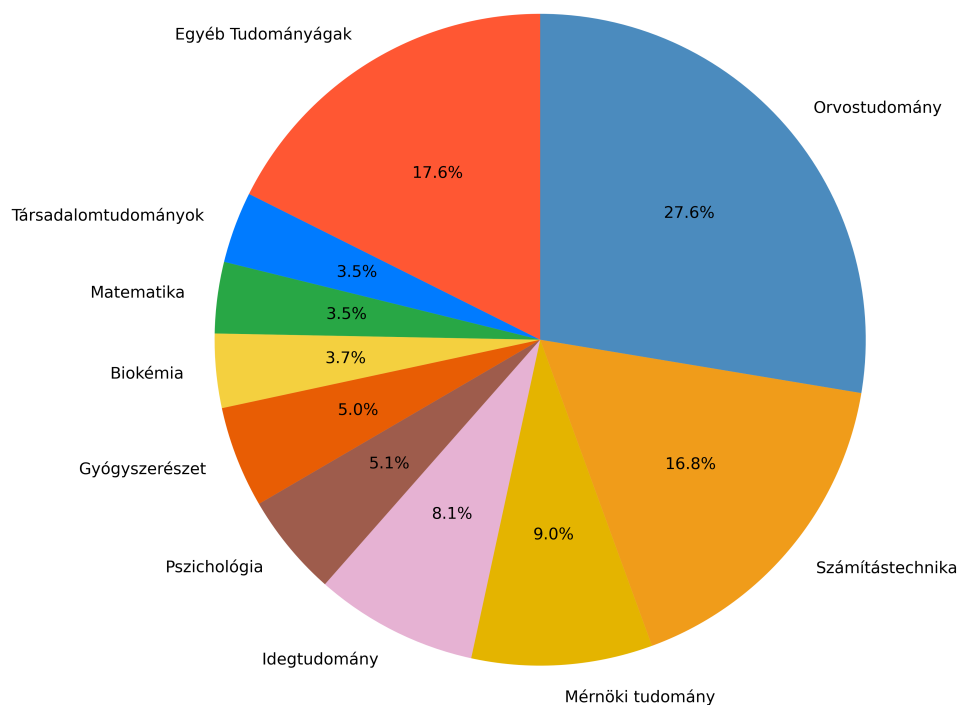
Számos tudományos munka készült már kifejezetten a szerencsejáték-függőségről. Bizonyos kutatások [8] kifejtik, hogy a terület növekedésében kiemelhető szerepe van a technológiai fejlődéseknek, ugyanis jelentős előnnyel bírnak a kialakult online platformok: hozzáférhetőség, megfizethetőség, anonimitás, kényelem és a menekülés lehetősége. A [6] cikk vizsgálta a hozzá köthető hátrányokat, ugyanis ezen platformok elérhetősége, az interaktív felület és a pénz könnyű költsége aggodalmakat vet fel a rendellenes szerencsejáték előfordulásának növekedésével kapcsolatban.



1.1. ÁBRA. Megjelent szakirodalmak száma évi bontásban

A kutatási téma tudományterületeken átívelő (lásd 1.2. ábra), a szerencsejáték-függőség mellett gépi tanuló algoritmusokat használtak más szenvedélybetegségek prediktálására is. Tudományos kutatásomhoz hasonlóan készült már gépi tanuló algoritmusokra épülő modell az alkoholhasználati zavar vizsgálatára, amely során klaszterezték a kutatásban [13] résztvevőket. Nem csak fizikai, hanem a mentális betegségek előrejelzésére is készültek gépi tanulásra épülő modellek, amelyet a [15]-ös szakirodalom fejt ki.

Ezen kutatások megvalósítása gépi tanuló algoritmusokkal meglehetősen sokszínű. Természetes nyelvfeldolgozást (NLP) alkalmazva több, mint 800 ezer közösségi média posztot vizsgáltak egy kutatásban [10], melyek segítségével olyan modell jött létre, ami képes előrejelezni a mentális betegségeket. Egy orvostudományi kutatás [12] eredményeképpen támasztó vektor (SVM) alapú osztályozó modellek segítségével már képesek prediktálni a tudatmódosító szer hatása alatt lévő embereket az agyi röntgenképek alapján. Megerősítési tanulási rendszer segítségével hoztak létre tudományos cikket [9] és modellt az internetfüggőség észlelése érdekében, amely oktatási intézményeknek és klinikáknak segít a kockázatok azonnali felismerése érdekében.



1.2. ÁBRA. Megjelent szakirodalmak tudományterületek szerint

A szakirodalmak feldolgozása során látható, hogy az adott kutatási terület nagy jelentőséggel és lehetőséggel bír. Számos kutatás készült a szerencsejáték-függőség súlyosságáról a különböző technológiai fejlődések eredményeképpen, illetve más szenvedélybetegségek felismerésére már több gépi tanuló algoritmusokra épülő modell készült. Azonban a szakirodalmi áttekintés során nem találtam olyan kutatást, amely pontosan a szerencsejáték-függőség előrejelzésére fókuszál, ez növeli a tudományos munkám fontosságát, amely motivál a kutatásom elkészítésében és egy ipari gyakorlatban alkalmazható modell létrehozásában.

2. Paraméterek meghatározása

A szerencsejáték-függőség határaival kapcsolatban nincs egységes szakmai definíció, így a hozzá köthető paraméterek sem egységesek, a szakirodalmak eltérnek a befolyásoló paraméterek, illetve azok súlyának tekintetében.

2.1. Paraméterek meghatározása NLP segítségével

A *Natural Language Processing* felhasználásával napjainkban már jelentős mennyiségű adatot lehet feldolgozni rövid idő alatt. Tudományos munkám során a ChatGPT modellt kértem meg, hogy a csatolt szakirodalom alapján határozza meg a függőséget definiálni paramétereket.

Prompt:

Kedves ChatGPT! TDK dolgozatom elkészítéséhez egy olyan ML klaszterező modellt szeretnék készíteni, amely fogadási adatok alapján meghatározza egy ügyfélről, hogy sportfogadási szenvedélybeteg vagy sem. A csatolt dokumentum alapján kérlek határozd meg azokat a paramétereket, amelyek relevánsak lehetnek a modell felépítésében. Kérlek vedd figyelembe, hogy csak a fogadási adatok (fogadás ideje, tét, nyeremény) állnak rendelkezésemre, amelyekből tetszőleges származtatott paramétert tudok készíteni.

Csatolt szakirodalom: *Current Addiction in Youth: Online Sports Betting* [1]

Válasz:

A dokumentum alapján az alábbi paraméterek relevánsak lehetnek a sportfogadási szenvedélybetegség meghatározásához az ML klaszterező modellben:

- Fogadási időpontok gyakorisága:
 - Napi/heti fogadások száma
 - Időintervallum az egyes fogadások között
- Tétek és nyeremények:
 - Fogadások átlagos tétje
 - Fogadások átlagos nyereménye
 - Maximális és minimális tétek és nyeremények
 - Nyertes fogadások aránya
- Tét és nyeremény változásai:
 - Tétek és nyeremények időbeli növekedése/csökkenése
 - Nagy tétek/nyeremények gyakorisága
- Származtatott paraméterek:
 - Tétek aránya az összes rendelkezésre álló tőkéhez képest
 - Fogadási aktivitás időbeli eloszlása (pl. hétvégi, éjszakai fogadások)

Ezek a paraméterek segíthetnek azonosítani a potenciális sportfogadási szenvedélybetegteket a fogadási adatok alapján.

Prompt:

Köszönöm! A következő cikk segítségével határozz meg újabb paramétereket, amelyek relevánsak lehetnek!

Csatolt szakirodalom: *Problem gambling worldwide: An update and systematic review of empirical research* [4]

Válasz:

Az általad megadott cikkek alapján, az alábbi paraméterek relevánsak lehetnek egy sportfogadási szenvedélybetegség detektálására szolgáló ML klaszterező modell felépítésében:

- Fogadási gyakoriság: Az időegységenként (pl. nap, hét) megtett fogadások száma.
- Tét nagysága: Az egyes fogadásokon elköltött összegek átlaga és szórása.
- Nyeremény gyakoriság és nagyság: Milyen gyakran nyer, illetve mekkora összeget nyer egy fogadó.
- Fogadási időpontok: Az időpontok, amikor a fogadásokat teszi (pl. éjszaka vs. nappal).
- Profitabilitás: Az összes nyeremény és az összes tét aránya.
- Veszteség mértéke: A veszteségek nagysága és gyakorisága.
- Hosszan tartó veszteségsorozatok: Az egymást követő veszteségek száma.
- Fogadási típusok változatossága: Különböző sportágakra vagy eseményekre tett fogadások sokfélesége.
- Tét növelése: A tét nagyságának növelése a veszteségek visszanyerése érdekében.
- Fogadási szünetek: Azok az időszakok, amikor a fogadó nem tesz fogadást.

Ezek a paraméterek segíthetnek a modellnek abban, hogy felismerje a problémás fogadási viselkedést, melyek közül sokra utal a sportfogadási szenvedélybetegség irodalma.

2.2. Adathalmaz bemutatása

A létrehozott adathalmaz közel 2 millió fogadási adatra épül, 18 oszlopot és az ügyfélazonosítót tartalmazza. Minden egyes sor egy ügyfél adatait jelöli, a 18 attribútumot az NLP modell tanácsai, saját tapasztalataim, és szakirodalmi kutatásomra alapozva hoztam létre. Az adatokat Oracle SQL és Python segítségével dolgoztam fel. Ezen attribútumok egyenként nem tartalmaznak érdemi információt a szenvedélybetegség lehetőségéről, azonban ezeket kombinálva és súlyozva már olyan modell hozható létre, amely képes lehet a predikcióra. Az alábbi fejezetben minden paramétert és létrehozásának lépéseit bemutatom.

2.2.1. Árbevétel

Az ügyfél sportfogadásra költött összege a vizsgált időintervallumban. Minden ügyfél esetében eltér ez az intervallum, ugyanis az utolsó fogadástól számoltam az intervallumot, így relevánsabb információt kaphatok a viselkedésről. Minden fogadott összeg árbevételt termel. Csak az az ügyfél kerülhet az adathalmazba, akinek legalább egy fogadása volt, így ez az érték minden ügyfél esetében pozitív és nincs hiányzó érték.

2.2.2. TJB

Az egyik legfontosabb sportfogadási paraméter a TJB (Tiszta Játék Bevétel), ami az árbevételek és nyereségek különbségét mutatja meg. A 2.1. táblázat példákon keresztül mutatja be a lehetséges TJB értékeket. Jelen esetben ugyanazon időintervallumra aggregáltam ezt az értéket, ami lehet negatív és pozitív érték is, hiányzó értéket nem tartalmazhat.

Árbevétel	Nyeremény	TJB
1000	0	1000
1000	5000	-4000
1000	300	700

2.1. TÁBLÁZAT. Példák TJB értékekre

2.2.3. Szelvénytér

A vizsgált időszakban lévő fogadások (szelvények) száma. Értéke csak pozitív egész lehet és nem tartalmazhat hiányzó értékeket.

2.2.4. Aktív napok száma

Megmutatja, hány különböző napon játszott az ügyfél az intervallumban. Az egyik legfontosabb attribútum lehet, ugyanis a rendszeres fogadás fontos tényezője a szenvedélybetegségnek. Értéke szintén csak pozitív egész lehet és nem tartalmazhat hiányzó értékeket.

2.2.5. Két fogadás között átlagos eltelt idő

A paraméter azt vizsgálja, hogy az ügyfél két fogadása között átlagosan mennyi idő telik el percben. Jól rámutat arra, hogy egy ügyfél milyen rendszerességgel játszik, a gyakori fogadás pedig fontos lehet a szenvedélybetegség kialakulásában. Azonban becsapós lehet: például ha egy ügyfélnek csak két fogadása van, akkor ezen két fogadás közötti időt mutatja. A többi paraméterhez hasonlóan ez sem releváns önmagában,

azonban más paraméterekkel együtt már hasznos információt ad. Értéke pozitív szám, megjelennek a hiányzó értékek azoknál az ügyfeleknél, akiknek csak egy fogadása volt a vizsgált időszakban.

2.2.6. Kiugró nyeresemény/veszteség utáni fogadás változás

Az *Interquartile Range* statisztikai mutató segítségével meghatározom egy ügyfél kiugró nyereseményeit és veszteségeit. Ezen fogadások időpontjától számított 3 napot tekintem befolyásolt időszaknak. A napokat súlyozva veszem figyelembe, tehát ha egy ügyfélnek egy nap két kiugró nyereseménye/vesztesége van, akkor ugyanazon napok kétszeresen vannak számolva a befolyásolt időszaknak. Ezután meghatározom minden ügyfélnek minden napra az átlagos árbevétel és TJB értékét, majd kiszámolom a befolyásolt időszakok és a teljes időszak átlagos értékeit. A két időszak mutatóiból arányszámot generálok. Ha az érték egynél nagyobb, akkor az ügyfél a kiugró nyeresemény/veszteség után nagyobb értékben fogadott vagy veszített. Egynél kisebb érték esetén az ügyfél fogadásai visszaesnek, ami csökkenti a függőségre való hajlam lehetőségét. A megjelenő hiányzó értékek azt jelentik, hogy az ügyfélnek nem voltak kiugró nyereseményei vagy veszteségei.

2.2.7. Kommunikáció utáni fogadási szokás változás

Az ügyfelek rendszeresen kapnak különböző kommunikációs megkeresést, például bónuszajánlatot, rendszerüzenetet vagy hírlevelet. Ezen megkeresések motiválhatják és emlékeztetik az ügyfeleket a sportfogadásról. A kommunikációkkal történő befolyásolhatóság szintén fontos lehet az addiktivitás kialakulásában. A meghatározáshoz ugyanazon függvényeket használtam: a kommunikációs megkeresést követő 3 nap viselkedését hasonlítottam össze a vizsgált időszak átlagos viselkedésével, majd a kettő különbsége adja meg a fogadási szokás változását árbevétel és TJB paraméterekkel.

2.2.8. Heti átlagok megoszlása

A heti átlagok megoszlása vizsgálja a rendszerességet, a játékos fogadásainak stagnálását. Több hétre vizsgálva határozom meg, hogy az adott ügyfélnek az adott héten milyen árbevétel és TJB értéke van, majd ezen heti aggregációkra határozom meg a szórást. Ez a statisztikai mutató jól mutatja, hogy milyen szintű rendszeresség jellemzi az adott ügyfelet két attribútum tekintetében.

2.2.9. Fogadási mutatók változása a regisztrációhoz képest

A szakirodalom kifejtették, hogy függőségre utaló jel lehet, ha egy ügyfél aktivitása folyamatosan növekszik. Ehhez az ügyfelek elmúlt 30 napját hasonlítom össze a regisztrációt követő első 30 naphoz. A két átlagos értéket beosztva arányszámokat

határozok meg. Jelentős mennyiségű hiányzó érték áll fenn, ugyanis az ügyfelek többségének nincs fogadása a regisztrációt követő első 30 napban.

2.2.10. Fogadási szünetek

A rendszerességet már több oldalról megközelítettem, azonban az NLP modell tanácsára vizsgálom, hogy az időintervallumban hány nap volt a maximális szünet, tehát azon egybefüggő napok száma, amikor az ügyfél nem fogadott. Azon ügyfelek rendelkeznek hiányzó értékekkel, akiknek csak 1 aktív napjuk volt a vizsgált időszakban.

2.2.11. Hosszan tartó veszteségsorozatok

A folyamatos veszteséget nehéz kezelni, bizonyos ügyfelek feladják a fogadásokat, többen pedig beleesnek ilyen sorozatokban, és veszteségek után még többet fogadnak. Az attribútum vizsgálja, hogy egyhuzamban mekkora volt a maximális veszteséges fogadások száma. Hiányzó értékek nincsenek, 0 értéket vesznek fel azon játékosok, akiknek nem volt vesztes fogadása.

2.3. Adattisztítás

A létrehozott adathalmaz számos hiányzó értéket tartalmaz, amelyeket kezelni szükséges. Ezen értékek tisztítására számos módszer létezik, például átlaggal való helyettesítés vagy regressziós modell építése a hiányzó adatokra. Tudományos munkám során nem tudtam ezen megszokott módszerekre építeni a hiányzó adatok mennyisége miatt. Szerencsére a legfontosabb mutatóknál (árbevétel, TJB, aktivitás) nincsenek hiányzó adatok, főként a különböző arányszámok esetében jelennek meg a hiányzó értékek. Jelen adathalmaz kidolgozásában a hiányzó értékek valamilyen jelentése van, a hiányzó értékek nem adatgyűjtési hibák következménye. Minden paraméter esetében feldolgoztam a hiányzó érték jelentését és a meglévő adatok értékkészletéhez viszonyított helyettesítését adtam meg:

- Két fogadás között eltelt idő esetén hiányzó érték jelenik meg, ha a vizsgált időszakban csak 1 db fogadása volt az ügyfélnek. Ezen hiányzó értékeket a maximális érték 110%-val helyettesítettem. A maximális érték a két darab, nagy különbséggel tett fogadással rendelkezők esetében volt. Hasonló módon jártam el a fogadási szünetek esetében hiányzó értékkel rendelkezők esetében.
- Kiugró események és kommunikáció utáni változásoknál a hiányzó értékek a nem befolyásolt ügyfeleket jellemzik, így ezen értékeket 0-val helyettesítettem.
- A regisztráció első időszakában nem fogadó ügyfelek hiányzó értékekkel rendelkeznek, amit 2-vel helyettesítettem és a fogadási gyakoriság dupláját jelöli.

3. Modell létrehozása

A módszer kialakításának első lépése a dimenziócsökkentő eljárások megvalósítása. A létrejövő koordináták csökkentik a nagy dimenziószámból adódó távolságot és segítik a vizualizációt. A dimenziócsökkentést megvalósítom két ismert eljárással és egy saját megközelítéssel. Minden eljárásra egy kezdeti klaszterezést hajtok végre, amely segítségével három különböző, függőket tartalmazó csoportot hozok létre. A modellezés hatékonyságát erősíti, ha egy problémára több megközelítés ad részmegoldást és a végeredmény ezen részmegoldások összedolgozásával születik.

A létrehozott kezdeti klasztereket ezután átdolgozom távolság mutatókra, amely három megközelítés szerint jelzi, hogy az adott ügyfél milyen távol van a függőség veszélyétől. Ezen távolság értékeket feldolgozva hozom létre a végleges, *k-Means* klaszterező eljárást.

Végeredményként egy bináris címkével tudom jelölni az ügyfeleket, amely jelzi, hogy a modell alapján függő vagy sem. Ezt a címkét felhasználva egy bináris osztályozó modellel ellenőrzöm, hogy egy új ügyfelet milyen pontossággal lehet osztályozni, ami fontos lehet a modell felhasználása során.

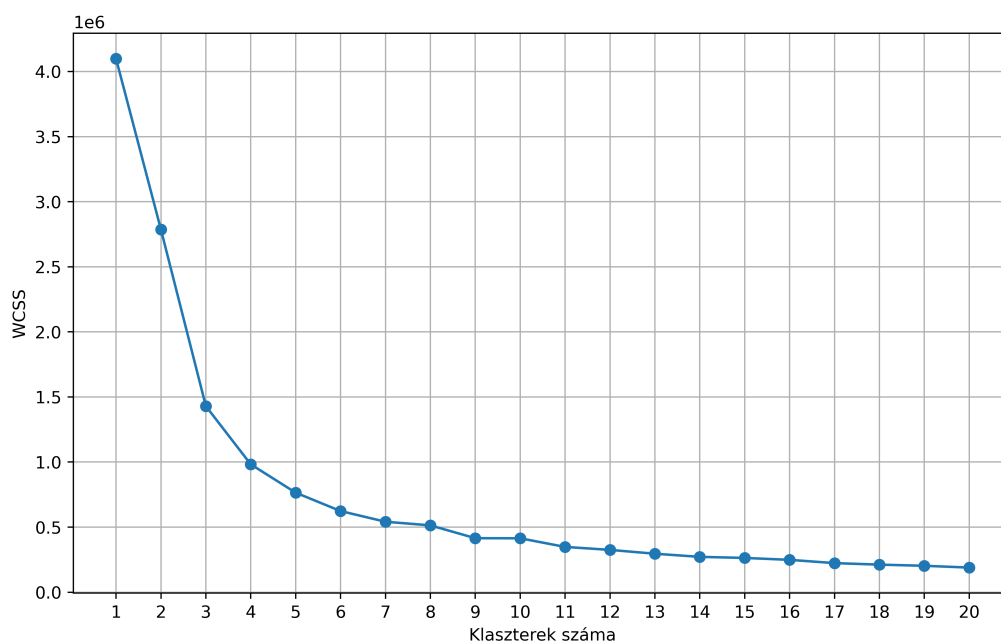
3.1. Dimenzicsökkentő eljárások

3.1.1. Uniform Manifold Approximation and Projection (UMAP)

Az eljáráshoz köthető tudományos cikk [11] kifejti, hogy az UMAP egy újfajta manifold-tanulási technika a dimenziócsökkentésre. Az UMAP egy olyan elméleti keretből épül fel, amely a Riemann-geometrián és az algebrai topológián alapul. Az eredmény egy gyakorlati, méretezhető algoritmus, amely valós adatokra alkalmazható. Az algoritmus jól megőrzi a globális struktúrát és gyors futási teljesítményt nyújt. Ezenkívül nincs számítási korlátja a beágyazási dimenziókra vonatkozóan, így általános célú dimenziócsökkentő technikaként is alkalmazható a gépi tanulásban.

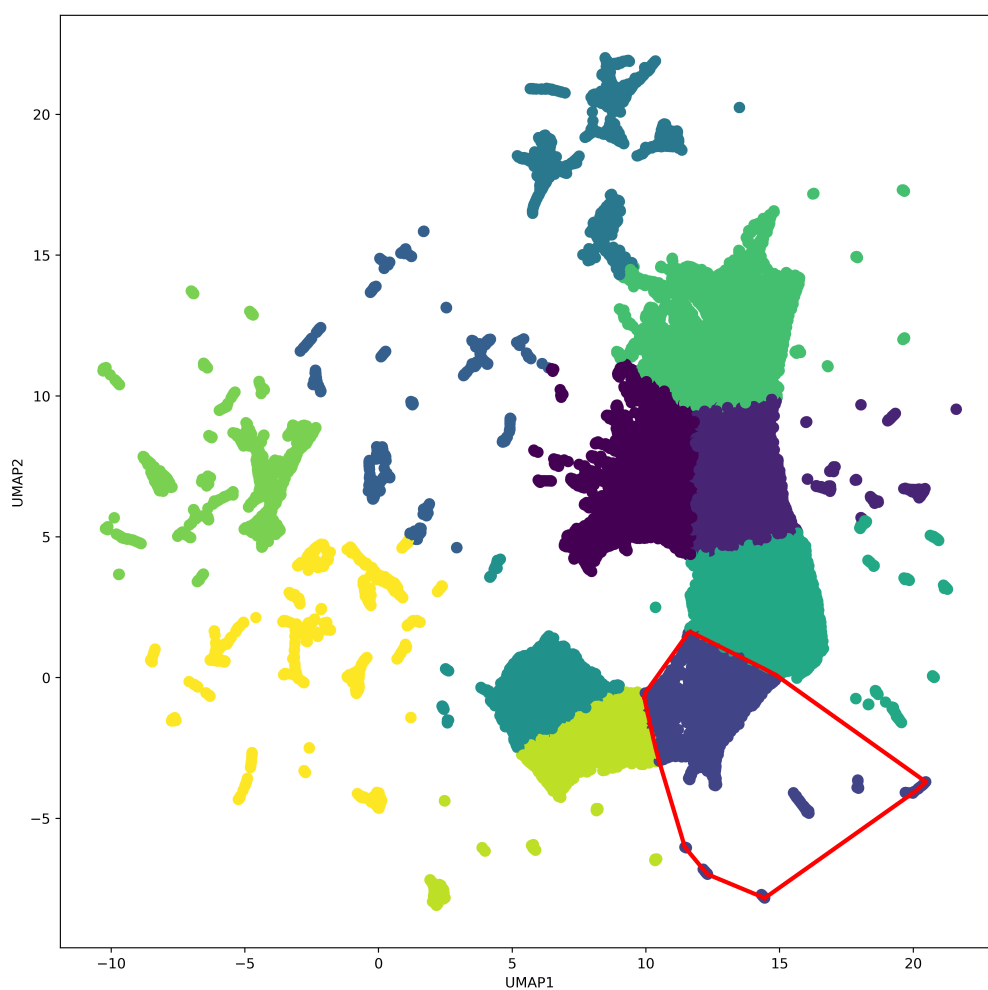
Az algoritmus alkalmazásának előfeltétele a standardizálás, amit a *StandardScaler()* eljárással valósítok meg. A redukált értékeket már fel tudja használni az UMAP eljárás, amit az *umap-learn* könyvtárral valósítok meg. Bemenő feltételnek megadom, hogy 15 szomszédot vegyen figyelembe a lokális szerkezet feltérképezésére és végeredményként kétdimenziós koordinátát adjon vissza. Ezek az általános és elterjedt feltételek az UMAP alkalmazásakor.

A kezdeti klaszterezés fontos eleme, hogy a klaszterek száma meghatározásra kerüljön, ehhez a könyökmódszert használtam. Jelen lépésnél fontosnak tartom, hogy nagy számú klaszter kerüljön meghatározásra, ugyanis így a függőket jellemző klaszterben kisebb lesz az elemszám, a többi klaszterben lévők a későbbiekben pedig jobban lehet differenciálni. A könyökmódszert felhasználva 11 klasztert hozok létre, mert ott érzékelhetőbb még egy kisebb ugrás (lásd 3.1. ábra), utána viszont az esés mértéke egyenletesebb, tehát már kevésbé lesznek homogénebbek a klaszterek.



3.1. ÁBRA. Könyökmódszer a kezdeti klaszterezésre

A kétdimenziós koordinátákra végrehajtottam a *k-Means* klaszterező eljárást 11 klaszterre. A létrejött klaszterezés után visszafejtettem az ügyfelek fogadási adatait átlagolva csoportonként. A 2-es számú klaszter az, amelyik a leginkább kiemelkedik (lásd 3.2. ábra) a többi klaszter közül és ők tekinthetők ideiglenes, függő klaszternek.



3.2. ÁBRA. UMAP klaszterek, kiemelve a 2-es számú klaszter

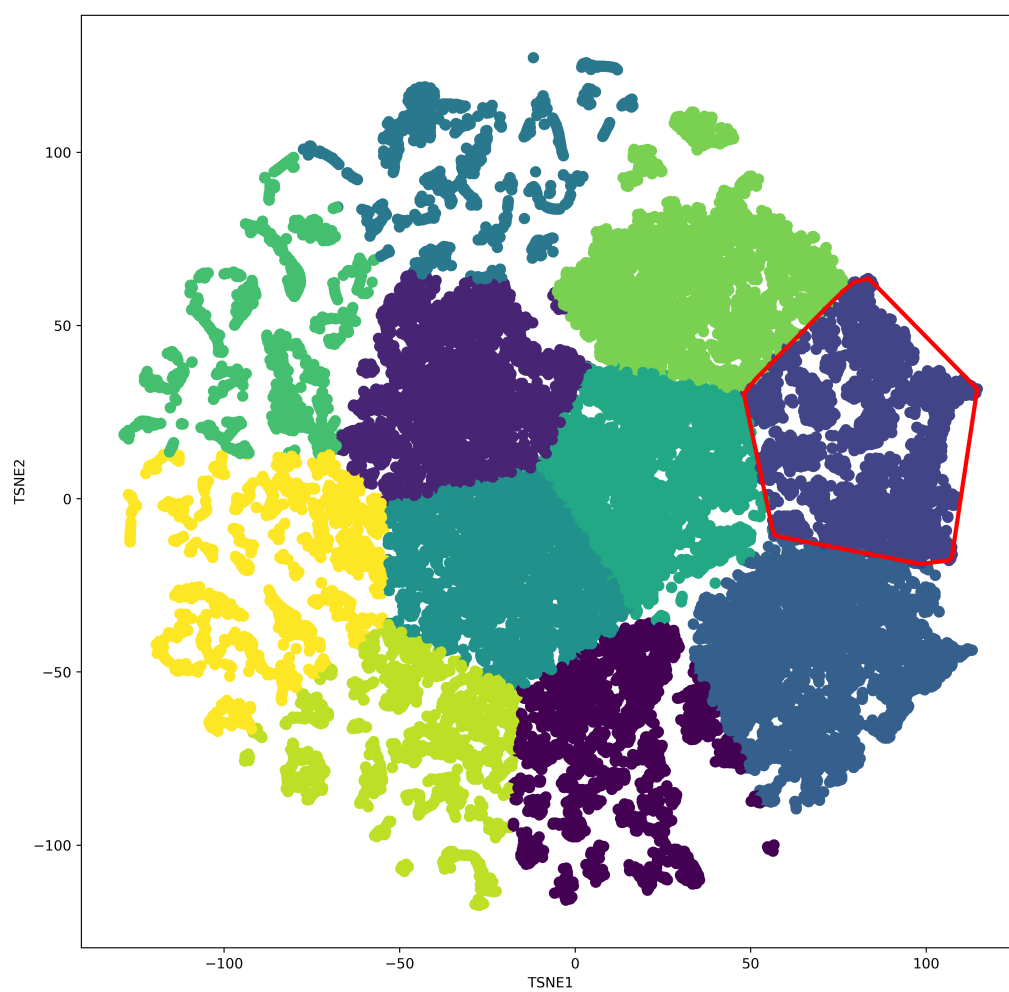
3.1.2. t-Distributed Stochastic Neighbor Embedding

A t-SNE egy nem-lineáris dimenziócsökkentő technika, amely az UMAP-hoz hasonlóan magas dimenziójú adathalmazokat két- vagy háromdimenziós térbe vetít le. A tudományos leírás [16] szerint fő előnye, hogy jobban megőrzi a lokális szerkezeteket, tehát a magas dimenzióban közel lévő adatpontok a redukált térben is közel lesznek egymáshoz. Hátránya azonban, hogy nagy adathalmaz esetén lassabb: tudományos munkám során a t-SNE modell futási ideje közel hatszorosa volt az UMAP futási idejének.

Klaszterek számának meghatározásához hasonlóan a könyökmódszert használtam, de ezen eljárásnál nem tapasztalható magasabb klaszterek esetében olyan "könyök", amely segítené a meghatározást, így az egységesség miatt az UMAP eljárással megegyező, 11 db klasztert használok.

A standardizálás után alkalmazom a *scikit-learn* könyvtár *TSNE* eljárását, ahol csak a dimenziószám megadása szükséges. A létrejött klasztereket visszavetítve és átlagolva a fogadási paraméterekre látható, hogy ezen klaszterezés esetében is kiemelhető egy klaszter. A dimenziócsökkentés után létrejött kétdimenziós tér a 3.3. ábrán látható, kiemelve a függőket tartalmazó klasztert.

Összegezve a két kezdeti klasztert, a legtöbb paraméterben hasonló átlagok találhatók és a játékoszám is közel megegyezik. A legnagyobb eltérés a két fogadás között eltelt időben figyelhető meg: az UMAP átlaga közel ötszöröse a t-SNE átlagának. Sőt, a t-SNE esetében nem is a függőknek definiált kezdeti klaszterben a legkisebb ez az érték. Ezen paraméterek is jól mutatják, hogy több eljárás összedolgozásával érhető el a valódi függőket tartalmazó klaszter, illetve a függőséget nem lehet azonosítani egyetlen paraméterrel.

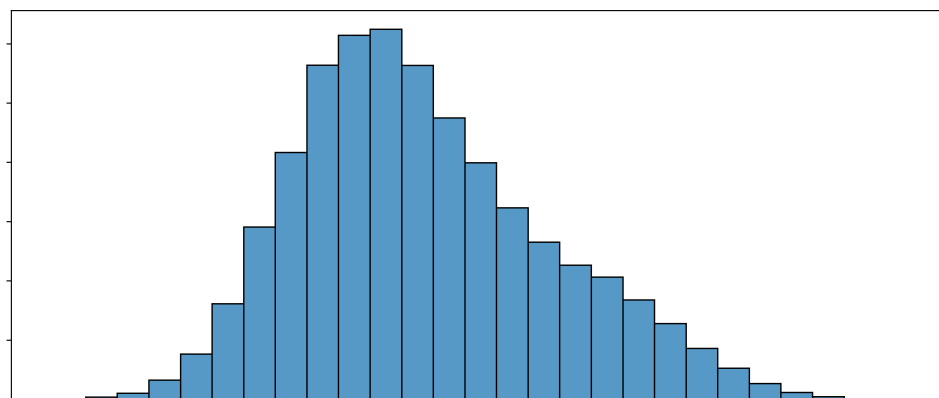


3.3. ÁBRA. t-SNE klaszterek, kiemelve a 2-es számú klaszter

3.1.3. Saját megközelítés

A létrehozott megközelítés során célom egy olyan numerikus jellemző létrehozása, amely reprezentálni tudja a függőségre való hajlamot a meglévő paraméterek felhasználásával. Az eljárás alapja, hogy minden numerikus paraméterhez megállapítok egy határt, amelyet átlépve az adott ügyfél az adott paraméter szerint függőnek tekinthető. Ezekre a határookra a feldolgozott szakirodalom nem ad javaslatot, így saját tapasztalataimra építve határozom meg az értékeket odafigyelve arra, hogy kellően szigorú határok legyenek, amelyek ténylegesen a függő ügyfeleket tudják jelölni.

Minden ügyfélre vizsgálom a paraméterhatárok teljesülését, és összegzem a teljesülő függőségi feltételek számát, így minden ügyfélre létrejön egy 0-18 közötti egész érték, amely jelzi a függőségre való hajlamot. Ezek az értékek balra eltolódó normális eloszlással jellemezhetők (lásd 3.4. ábra). Az ábrán az értékeket - az eredmény megtartásának figyelembevételével - torzítom az érzékeny adatok megőrzése érdekében.



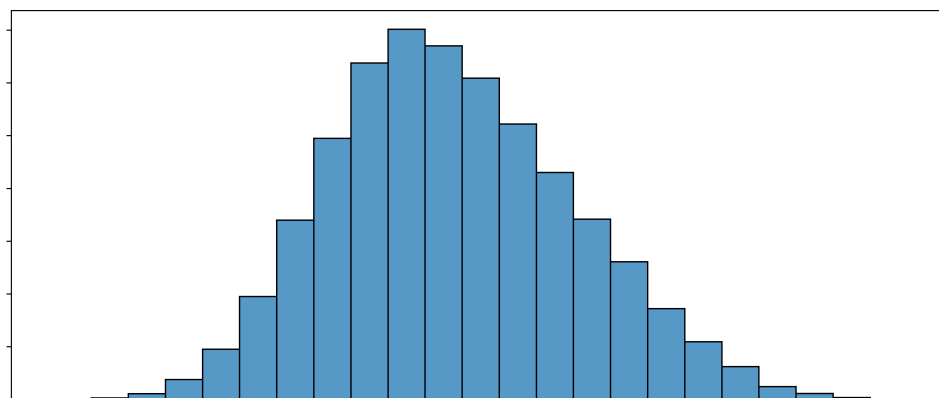
3.4. ÁBRA. Torzított függőségre utaló hajlam eloszlásdiagramja

A meglévő paramétereket vizsgálva látható, hogy egy részük összekapcsolható, illetve nem annyira relevánsak egyesével, hogy egy teljes értéként reprezentálják a függőségre utaló hajlamot, ezért az alábbi paramétereket összekapcsolom és egy új, összegzett paraméterként reprezentálom:

- Kiugró események utáni változások
 - Kiugró nyeresemény utáni árbevétel változás
 - Kiugró nyeresemény utáni TJB változás
 - Kiugró veszteség utáni árbevétel változás
 - Kiugró veszteség utáni TJB változás

- Kommunikációs megkeresés utáni változások
 - Kommunikáció megkeresés utáni árbevétel változás
 - Kommunikációs megkeresés utáni TJB változás
- Regisztrációhoz viszonyított változások
 - Vizsgált intervallum árbevétel változása a regisztrációs időszakhoz viszonyítva
 - Vizsgált intervallum TJB változása a regisztrációs időszakhoz viszonyítva

Az újonnan létrehozott paraméterek akkor felelnek meg a függőségi feltételeknek, ha legalább 1 integrálódó paraméter megfelel, tehát a paraméterek között "vagy" kapcsolat van. Az átalakítás után már lényegesebben homogénabbak és egyenértékűbbek lesznek a jellemzők, a 18 kiinduló paraméter helyett 12 paraméter maradt, így az ügyfelek 0-12 közötti értéket vehetnek fel a teljesült függőségi mutatók alapján. Egy ügyfél jelen megközelítés szerint akkor tekinthető függőnek, ha a létrejött jellemzők közül legalább 9 mutató esetén megfelel a függőségi kritériumnak. A redukált paraméterek esetén is fennmarad a balra eltolódó normális eloszlás, amelyet a 3.5. ábrán szintén torítottam az érzékeny adatok megőrzésének érdekében.



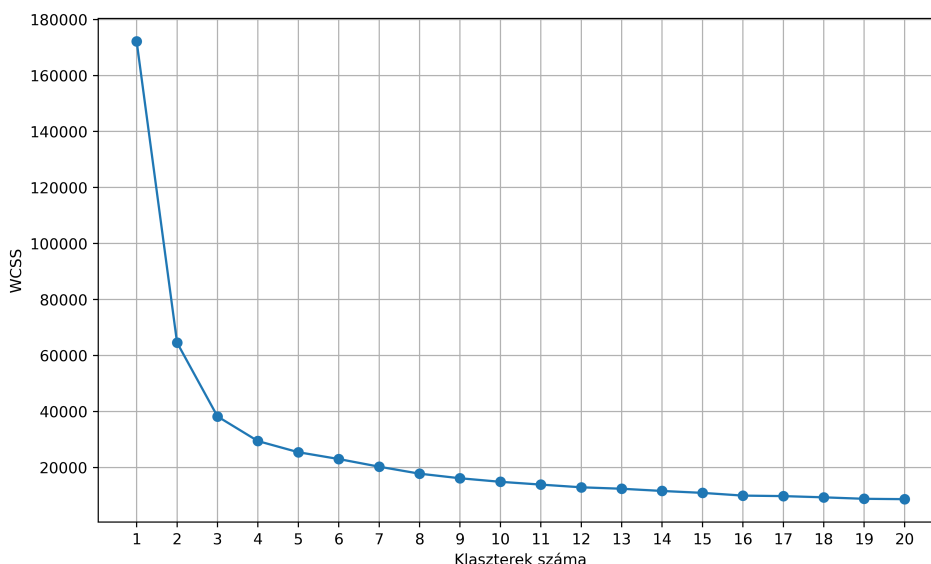
3.5. ÁBRA. Torzított függőségre utaló hajlam paraméterredukció után

3.2. Végleges klaszterek meghatározása

Az előző fejezetben három megközelítéssel hoztam létre kezdeti klasztereket a függőség prediktálására. Ezen megközelítések már relevánsan mutatják az ügyfelek függőségre való veszélyeztetettségét, azonban a megközelítések egybedolgozásával sokkal hatékonyabb és homogénabb végső játékoscsoportokat lehet létrehozni.

A végső megoldás a függő ügyfeleket tartalmazó kezdeti klaszterektől való távolságokon alapul. Első lépésként az UMAP és t-SNE eljárások által létrehozott kétdimenziós koordinátákat kell átdolgoznom. A klasztertől való távolság meghatározására euklideszi távolságot használok. Azonban nem a klaszterközepétől való távolságot keresem, hanem a minimális távolságra lévő pontot, és ezen pontok közötti távolsággal jellemzem a függőségtől való távolságot. Ennek az az oka, hogy a cél nem a függőség középpontjától való távolság meghatározása (ami nem is lehetséges), hanem az a távolság, amely jelöli a vizsgált pont függőnek tekinthetőségétől való távolságát. A kezdeti klaszterezés során függőnek jelölt ügyfelek 0 távolságértéket kapnak, hiszen ők nem jellemezhetőek függőségtől való távolsággal.

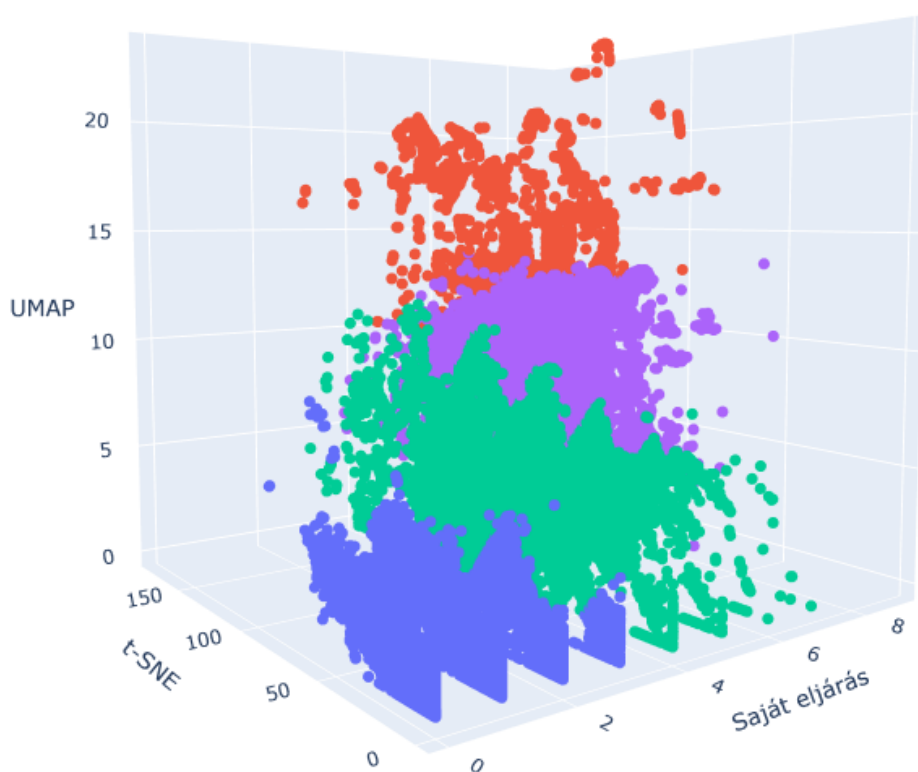
Saját megközelítem esetén lényegesen egyszerűbb a távolság meghatározása: függőnek tekinthető az az ügyfél, akinek legalább 9 paraméterre meghatározott feltétele teljesült, ők 0 távolságra vannak a függőségtől. A nem függők távolsága a teljesült paraméterek száma kivonva a bekerülési (9 paraméter) küszöbtől, így ebben a megközelítésben az ügyfeleket egy 0-9 közötti távolság jellemzi.



3.6. ÁBRA. Könyökmódszer a végső klaszterszám meghatározásához

A meghatározott paraméterek lényegesen eltérő értékkészlettel rendelkeznek, így standardizálni szükséges őket a klaszterezést megelőzően. A klaszterszám meghatározásához ismét a klaszterkönyök módszert használok, amely az esetleges klaszterek homogenitását vizsgálja (lásd 3.6. ábra). Az ábrát elemezve 4 db klasztert választok ki, amelyek közül 1 klaszter fogja a függő játékosokat tartalmazni. Ez természetesen eltérhet az ipari gyakorlatban, ugyanis a klaszterek száma erősen befolyásolja a függőnek tekintett játékosok számát.

A végleges klaszterezést az előzőleg bemutatott módszer alapján 4 klaszterre hajtom végre a *k-Means* eljárás segítségével. A létrejött klaszterek jól megfigyelhetők a 3.7. ábrán, amely mind a három távolságértéket tartalmazza.



3.7. ÁBRA. Távolság alapú klaszterezés eredménye

A kék színnel jelölt ügyfelek tekinthetők függőknek, ugyanis az ő távolságértékeik a legkisebbek és a könyökmódszer alkalmazásával egy jól szeparált klaszterként tekinthetők. Ez a folyamat három különböző, kezdeti klaszterezési és dimenziócsökkentő eljárás eredményeit dolgozza össze a homogenitás és hatékonyság növelése érdekében.

A meghatározott klaszterek alapján visszavetítettem a kezdeti paraméterekre, amelyek aggregálása és átlagolása után megbizonyosodtam a klaszterezés eredményéről, a kék színű klaszterbe tartozó ügyfelek függőségre utaló numerikus jellemzői valóban eltérnek és jelentős mértékben kiugrónak a többi klaszterbe tartozó ügyfelek mutatóihoz viszonyítva. Ezen megállapításra és az előző vizualizációra hivatkozva a tudományos munkám eredményét, a szerencsejáték-függőség prediktálását sikeresnek tekintem.

3.3. Ellenőrzés felügyelt gépi tanuló algoritmusokkal

A létrejött klaszterezés során megállapításra került a függő klaszter, minden más klaszterbe tartozó adatpontokat pedig nem függőnek tekintek, így létrejött egy bináris felbontása az adathalmaznak. A valós alkalmazás során új adatpontokról szükséges eldönteni a függőséget a betanított modell alapján.

Az alkalmazhatóság validálása érdekében létrehozok egy bináris osztályozási modellt, amely osztálycímkei a klaszterezés során létrehozott *függő* vagy *nem függő* csoportok lehetnek. Ez egy kritikus pontja a tudományos munkámnak, ugyanis egy jól klaszterező eljárás értéke radikálisan csökkenhet, ha az új adatpontok osztályozásának pontossága alacsony.

Az osztályozáshoz a *scikit-learn* könyvtárat és a *Random Forest* eljárást alkalmazom, amely döntési fák egy csoportját használja fel az osztályozási (és regressziós) feladatok megoldására. A szerzők a [3]-ban igazolták, hogy több döntési fa segítségével javítja az előrejelzések pontosságát és csökkenti a túlilleszkedés kockázatát.

Az ellenőrző modell bemenő paraméterei a klaszterezés során is használt 18 numerikus jellemző, célváltozója a bináris, függőséget leíró jellemző. Az adathalmazt 80-20% arányban osztom fel tanító és teszt adathalmazokra. A pontosság mérése miatt a klaszterezett adathalmazt használom fel az ellenőrzésre.

Bináris osztályozási modell teljesítményének mérésére megfelelő mutató az *Accuracy Score*, amely megmutatja, hogy a tesztelés során létrehozott predikciók milyen arányban találta el a valódi osztályt. Az ellenőrző modell esetében ez a mutató 0.97, ami egy sikeres osztályozásra utal.

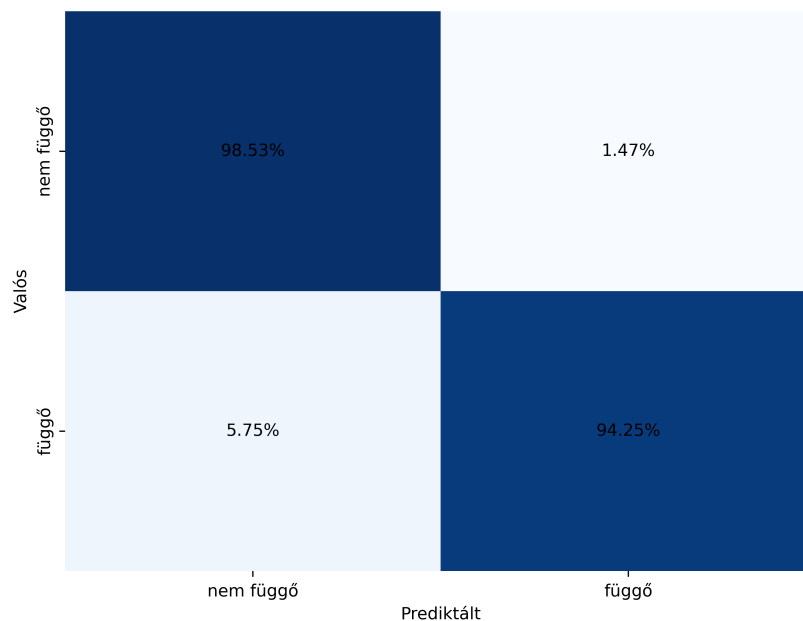
Azonban az adataim kiegyensúlyozatlanok, hiszen az ügyfelek jelentős többsége a nem függő osztályba tartozik, így megvizsgáltam más teljesítménymutató metrikákat: a *Precision* jelzi, hogy milyen pontosan találja el a modell az egyes osztályokat, a *Recall*

a valóban függő egyedek megtalálását jelzi, az *F1-score* pedig az előző két mutató harmonikus átlaga. Minden metrika kiértékelése mutatja, hogy az ellenőrzés sikeresnek tekinthető és megvalósítható az új adatpontok besorolása (lásd 3.2. táblázat).

Accuracy Score:	0.97
Precision:	0.96
Recall:	0.94
F1-score:	0.95

3.2. TÁBLÁZAT. Teljesítménymutató metrikák eredményei

A 3.8. ábrán található konfúziós mátrix megmutatja az osztályozott adapontokat a valós és prediktált címkék szerint. A sorok szerint százalékosított értékek mutatják, hogy a függő egyedek közel 6%-a nem függőnek lett előrejelezve (ami a Recall mutató is egyben). Úgy gondolom, hogy ez továbbfejlesztési területnek kijelölhető, hiszen célom ennek a mutatónak a minimalizálása.



3.8. ÁBRA. Konfúziós mátrix soronként százalékosítva

4. Elméleti modellezés

Jelen fejezet célja a létrehozott gépi tanuló algoritmusokra épülő informatikai modell megközelítése és reprezentálása matematikai szemszögből. Az elméleti modell valamilyen vizsgált rendszerben lejátszódó jelenség, folyamat vagy tevékenység a vizsgált szempontjából lényeges tulajdonságai közötti összefüggések megfogalmazása, amely a matematika szimbólumrendszerén keresztül teremt kapcsolatot a vizsgált rendszer be- és kimenő jellemzői között.

4.1. Dimenziócsökkentő eljárások modellezése

A dimenziócsökkentő eljárások iteratív optimalizációs modellt használnak, így nem lehet egzakt, zárt matematikai képletekkel kifejezni, azonban a módszerek alapjául szolgáló elvek felírhatók matematikai modellek segítségével.

4.1.1. t-SNE modellezése

A *t-Distributed Stochastic Neighbor Embedding (t-SNE)* eljárás a pontok közötti valószínűségeken alapul. Ahogyan a szerzők megmutatták a [16]-ban, első lépésként a magas dimenzióban lévő adatpontokat reprezentálja egy lokális valószínűséggel. Egy adatpont x_i és egy másik adatpont x_j közötti távolságot a következőképpen alakítja át egy valószínűségre:

$$p_{ij} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)}$$

Itt p_{ij} egy sztochasztikus valószínűség, amely azt fejezi ki, mennyire valószínű, hogy az x_i és x_j pontok egymás szomszédai.

Második lépésként a létrejött alacsony dimenziós térben az y_i és y_j pontok közötti valószínűséget határozza meg:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

A t-SNE célja, hogy minimalizálja a Kullback-Leibler divergenciát a p_{ij} és q_{ij} valószínűségek között. A Kullback–Leibler divergencia (KL divergencia) egy statisztikai mérőszám, amely két valószínűségi eloszlás közötti különbséget méri. Az [5] könyv szerint leginkább arra használják, hogy meghatározzák, mennyi információ veszik el, ha egy valódi eloszlást egy másik eloszlással közelítünk. Az optimalizálás legtöbbször gradiens módszerrel valósul meg.

$$D_{\text{KL}}(p_{ij} \parallel q_{ij}) = \sum_{i \neq j} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right)$$

4.1.2. UMAP modellezése

A szerzők leírása szerint [11] a *Uniform Manifold Approximation and Projection* eljárás geometriai megközelítést alkalmaz. Elsőként az adatpontok szomszédságával létrehoz egy környezeti gráfot, a szomszédsági struktúra meghatározásához egy exponenciális eloszlást használ:

$$w_{ij} = \exp(-r_i \|x_i - x_j\| - d_k(x_i))$$

Itt w_{ij} a kapcsolat (súly) az x_i és x_j pontok között, r_i a lokális skálázó paraméter az x_i ponthoz, $d_k(x_i)$ pedig az x_i pont k -adik szomszédjához mért távolság. A $\|x_i - x_j\|$ az x_i és x_j közötti euklideszi távolságot jelenti.

A grafikus struktúra létrehozása során minden pont körüli szomszédsági kapcsolatokat figyelembe kell venni, így a szomszédsági gráf élei a következőképpen alakulnak:

$$W_{ij} = \begin{cases} w_{ij} & \text{ha } x_j \text{ a } k\text{-adik legközelebbi szomszédja} \\ 0 & \text{minden más esetben} \end{cases}$$

A létrejött alacsony dimenziós térben az elhelyezett pontok közötti valószínűséget az előzőleg bemutatott t-SNE eljáráshoz hasonlóan határozza meg. A költségfüggvény minimalizálása szintén a gradiens módszerrel történik.

4.1.3. Saját megközelítés modellezése

A saját megközelítem egy egydimenziós térben adja vissza, hogy a definiált függőségi határok közül mennyi teljesül az adott esetre:

1. **Paraméterek Definíciója:** Legyen $X = \{x_1, x_2, \dots, x_{18}\}$ a bináris paraméterek halmaza, ahol $x_i \in \{0, 1\}$ minden i -re.
2. **Kapcsolódó Paraméterek:** Definiálok egy C halmazt, amely tartalmazza a kapcsolódó paraméterek csoportjait. Minden $C_j \subseteq X$ egy csoport, amely tartalmazza az összefüggő paramétereket.
3. **Összevonás:** Az új paraméterek, $Y = \{y_1, y_2, \dots, y_{12}\}$, a következőképpen definiálhatók:

$$y_j = \max_{x_i \in C_j} x_i \quad (j=1, 2, \dots, 12)$$

Ez azt jelenti, hogy ha a csoportban bármelyik paraméter értéke 1, akkor az összevont paraméter értéke is 1.

4. **Szummázás:** A 12 paraméter szummázása, hogy egy egydimenziós érték legyen a végeredmény:

$$S = \sum_{j=1}^{12} y_j$$

S az egydimenziós térben reprezentált érték, amely megmutatja az összevont függőségi határok teljesülésének mennyiségét. Minél nagyobb az S értéke, az adott esetre annál több függőségre utaló jel teljesül.

4.2. Távolságok modellezése

A *T-SNE* és *UMAP* eljárások segítségével kétdimenziósra redukáltam az eredetileg 18 paraméterből álló eseteket. Egy eljárás esetében a kiválasztott klasztertől való távolság az alábbi módszerrel határozható meg:

- Legyenek a pontok $P = \{p_1, p_2, \dots, p_n\}$, ahol minden p_i egy kétdimenziós koordinátájú pont
- Ha egy pont p_i benne van az adott C_j klaszterben, akkor a távolság 0:

$$d(p_i, C_j) = 0 \quad \text{ha} \quad p_i \in C_j$$

- Ha egy pont p_i nincs benne az adott C_j klaszterben, akkor a távolság az adott pont és a klaszterhez tartozó legközelebbi pont közötti euklideszi távolság:

$$d(p_i, C_j) = \min_{p_k \in C_j} \|p_i - p_k\|$$

ahol az euklideszi távolság $\|p_i - p_k\|$ két pont között a következő:

$$\|p_i - p_k\| = \sqrt{(x_i - x_k)^2 + (y_i - y_k)^2}$$

ahol $p_i = (x_i, y_i)$ és $p_k = (x_k, y_k)$ a pontok 2D koordinátái.

Tehát az adott pont p_i távolsága az j -edik klasztertől:

$$d(p_i, C_j) = \begin{cases} 0, & \text{ha } p_i \in C_j, \\ \min_{p_k \in C_j} \|p_i - p_k\|, & \text{ha } p_i \notin C_j. \end{cases}$$

A létrejött távolságok különböző értékészletten rendelkeznek, így azokat szükséges standardizálni azaz a távolságokból kivonni a μ átlagot, majd elosztani a σ szorással:

$$d'_{\text{t-SNE}}(p_i, C_j) = \frac{d_{\text{t-SNE}}(p_i, C_j) - \mu_{\text{t-SNE}}}{\sigma_{\text{t-SNE}}}$$

$$d'_{\text{UMAP}}(p_i, C_j) = \frac{d_{\text{UMAP}}(p_i, C_j) - \mu_{\text{UMAP}}}{\sigma_{\text{UMAP}}}$$

$$d'_{\text{saját}}(p_i) = \frac{d_{\text{saját}}(p_i) - \mu_{\text{saját}}}{\sigma_{\text{saját}}}$$

A standardizált távolságok segítségével klaszterezem a létrehozott adathalmazt és meghatározom a függőségben szenvedő ügyfelek csoportját.

5. Eredmények kiértékelése

A klaszterezés eredményeképpen a vizsgált adathalmazban lévő ügyfelek egy részét jelöli a modell szerencsejáték-függőnek. A valós, online sportfogadási adatokra épülő csoportosítás során elkülönített játékosok pontos elemszámát az adatok érzékenysége miatt nem publikálhatom, így összehasonlítást készítek más országok kutatásaival.

A Psychiatric Times folyóiratban megjelent cikk [2] alapján az Egyesült Államokban az American Psychiatric Association (APA) becslése szerint a felnőtt lakosság körülbelül 1%-a mutat szerencsejáték-függőségi tüneteket. Ez az arány alapvetően az APA által használt diagnosztikai kritériumokra épül.

Az Egyesült Királyságban a szerencsejáték-függőség előfordulását a Gambling Commission 2018-as jelentése alapján [7] a felnőttek 0.6%-ára becsülik. Ez az adat a nemzeti egészségügyi felméréseken alapul, amelyek a problémás szerencsejáték szintjeit vizsgálják az ország különböző részein.

A Productivity Commission [14] riportja szerint szerencsejáték problémával küzd az ausztrál felnőtt lakosság 0.5-1%-a, ami 80-160 ezer embert jelenthet (Productivity Commission, 2010).

A három ország által készített elemzések közel megegyeznek a százalékos arányokban. A kutatásom során kapott elemszámot arányosítom a magyarországi lakossággal, illetve figyelembe veszem, hogy az adott vállalat mekkora részét teszi ki az itthoni online sportfogadó tevékenységeknek. Az arányosítás után az a következtetés szűrhető le, hogy a külföldi tendenciák megfigyelhetők és a magyarországi lakosság közel 1%-a tekinthető szerencsejáték-függőnek.

6. Összefoglalás

Napjainkban a szerencsejáték-függőség előfordulása növekvő tendenciában van. Az internetes tér sokféle lehetőséget kínál ezen tevékenységek megvalósítására, ugyanis számos ösztönző funkciójával kényelmes szolgáltatást kínálnak a vállalatok. A gépi tanulás és mesterséges intelligencia elterjedésével azonban már olyan eljárások alakíthatóak ki, amelyek segíteni tudják a szenvedélybetegségek kialakulásának megelőzését.

Tudományos munkám célja egy olyan gépi tanuló algoritmusokra épülő modell létrehozása, amely fogadási adatokból származtatott, függőségi hajlamra utaló paraméterek alapján klaszterezi az ügyfeleket és meghatároz egy függő ügyfeleket tartalmazó csoportot. Szakirodalmi áttekintésben vizsgáltam a témához köthető kutatási eredményeket a szenvedélybetegségek prediktálásával kapcsolatban. Az a megállapítás vonható le, hogy a szerencsejáték-függőség vizsgálata egyre több embert érint, és az egyéb szenvedélybetegségekkel ellentétben a szerencsejáték-függőség prediktálásával kapcsolatban korábban sem nemzetközi, sem hazai szakirodalmak nem jelentek meg. Így kutatásom kontribúciója, hogy valós fogadási adatokra épülő adatmodell létrehozásával egy fontos, de korábban nem megoldott problémára teszek javaslatot.

A kutatás első lépéseként különböző szakirodalmakra alapozva, NLP modell segítségével állapítottam meg a függőségre utaló numerikus jellemzőket. A kidolgozott 18 paramétert az *Uniform Manifold Approximation and Projection* és *t-Distributed Stochastic Neighbor Embedding* dimenziócsökkentő eljárásokkal dolgoztam fel a vizualizálhatóság érdekében, majd egy saját megközelítést is alkalmaztam az alacsonyabb dimenzióban történő jellemzés érdekében. Mind a három eljárás eredményeképpen meghatároztam egy kezdeti, függő ügyfeleket tartalmazó csoportot.

A kiinduló eljárások eredményeképpen már három különböző megközelítéssel megvalósult a klaszterezés. A végső csoportosításhoz ezen részmegoldásokat dolgoztam fel, ugyanis a különböző eljárások felhasználásával egy sokkal homogénabb csoport jöhetett létre. A kezdeti eljárásokban történő elhelyezkedéseket euklideszi távolságokra alakítottam át, így minden ügyfélhez meghatározásra került egy 3 paraméterből álló standardizált távolságérték. Ezen adathalmaz felhasználásával hajtottam végre a klaszterezést a *k-Means* algoritmus felhasználásával és meghatároztam a végső, függő ügyfeleket tartalmazó csoportot. A modell által létrehozott címkéket egy felügyelt gépi tanulási algoritmus segítségével ellenőriztem, amely igazolta a modell felhasználhatóságát és kiterjesztését az ipari gyakorlatban. Végül a matematikai szimbólumrendszer használatával is bemutattam a kidolgozott modellt.

A sikeres klaszterezés ellenére számos továbbfejlesztési javaslatot fogalmaztam meg, amelyek közül kiemelhető a végső klaszterezési algoritmus finomhangolása és hibásan osztályozott esetek csökkentése. Ezen további kutatási tevékenységek megvalósítása kellően motiváló, hiszen ténylegesen egy súlyos és korábban nem kutatott témára tesz megoldási javaslatokat.

7. Irodalomjegyzék

- [1] Aragay, N., Pijuan, L., Cabestany, À., Ramos-Grille, I., Garrido, G., Vallès, V., & Jovell-Fernández, E. Current Addiction in Youth: Online Sports Betting. *Front Psychiatry* (2021), 11:590554.
- [2] Black, D. W., & Shaw, M. Psychiatric comorbidity associated with pathological gambling. *Psychiatric Times* (2008), 25(12).
- [3] Breiman, L. Random Forests. *Machine Learning* (2001), 45(1), pp 5-32.
- [4] Calado, F. & Griffiths, M. D. (2016). Problem gambling worldwide: An update and systematic review of empirical research . *J Behav Addict* (2000–2015), 5(4), 592–613.
- [5] Cover, T. M., & Thomas, J. A. Elements of Information Theory (2006).
- [6] Gainsbury, S. M. Online gambling addiction: The relationship between internet gambling and disordered gambling. *Journal of Behavioral Addictions* (2015).
- [7] Gambling Commission (UK) (2021). Gambling Behaviour Report. <https://www.gamblingcommission.gov.uk/about-us/guide/page/problem-and-at-risk-gambling> [Letöltve: 2024. Október]
- [8] Griffiths, M. D. Internet gambling: Issues, concerns, and recommendations. *CyberPsychology & Behavior* (2004), 6(6).
- [9] Ji, H.-M., Chen, L.-Y., & Hsiao, T.-C. Real-time detection of internet addiction using reinforcement learning system. In *GECCO '19: Proceedings of the Genetic and Evolutionary Computation Conference Companion* (2019), pp. 1280-1288.
- [10] Low, D. & Rumker L. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study *Journal of Medical Internet Research* (2020), 22(10).
- [11] McInnes, L., Healy, J., & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *The Journal of Open Source Software* (2018), 3(29), 861.

-
- [12] Mete, M., Sakoglu, U., Spence, J. S., Devous Sr., M. D., Harris, T. S., & Adinoff, B. Successful classification of cocaine dependence using brain imaging: a generalizable machine learning approach. *BMC Bioinformatics* (2016), 17(Suppl 13), Article 357.
- [13] Mumtaz, W., Vuong, P. L., Likun, X., Malik, A. S., & Abd Rashid, R. B. Automatic diagnosis of alcohol use disorder using EEG features. *Knowledge-Based Systems* (2016), 105, 48-59.
- [14] Productivity Commission (Australia) (2010). Gambling Report. <https://www.pc.gov.au/inquiries/completed/gambling-2010/report> [Letöltve: 2024. Október]
- [15] Srividya, M., Mohanavalli, S., & Bhalaji, N. Behavioral modeling for mental health using machine learning algorithms. *Journal of Medical Systems* (2018), 42, 88.
- [16] van der Maaten, L., & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* (2008), 9, 2579-2605.