BUDAPESTI GAZDASÁGI EGYETEM

PÉNZÜGYI ÉS SZÁMVITELI KAR

SZAKDOLGOZAT

Molnár Kitti
Nappali
Gazdaságinformatika
Üzleti adatelemző

2024

BUDAPESTI GAZDASÁGI EGYETEM
ALKALMAZOTT TUDOMÁNYOK EGYETEME

PÉNZÜGYI ÉS SZÁMVITELI KAR

BUDAPESTI GAZDASÁGI EGYETEM

PÉNZÜGYI ÉS SZÁMVITELI KAR

A mesterséges intelligencia elfogultsága és annak hatása a döntéshozatalra

Belső konzulens: Kovács Endre

Külső konzulens: Kuknyó Dániel

Molnár Kitti
Nappali
Gazdaságinformatika
Üzleti adatelemző

2024

1149 Budapest, Buzogány utca 10-12.
Telefon: (+36-1) 469-6600
Fax: (+36-1) 469-6610
www.uni-bge.hu

Tartalomjegyzék

1. Bevezetés

1.1. A mesterséges intelligencia elfogultságának lényegessége és időszerűsége

A mesterséges intelligencia (MI) az elmúlt években a sok fejlődésnek köszönhetően egyre elterjedteb lett a világban. A mindennapokat is befolyásolja nem csak a nagy fejlődések miatt, hanem egyre több tevékenységben és munkakörben is alkalmazni kezdték a mesterséges intelligenciát.

Az egyik legkritikusabb és egyben leginkább figyelmen kívül hagyott probléma az MI rendszerek elfogultsága. Az elfogultság az MI rendszerekben különböző formákban jelenhet meg, és sok esetben súlyos következményekkel járhat. Az MI elfogultsága nem csupán technikai kérdés, hanem mélyreható társadalmi és etikai vonatkozásokkal is bír, amelyek befolyásolják a technológia hitelességét, megbízhatóságát és igazságosságát.

Az emberi döntéshozatalban az elfogultság mindig is jelen volt, azonban a mesterséges intelligencia alkalmazása új dimenzióba helyezi ezt a kérdést. Az MI algoritmusok által hozott döntések gyakran szubjektív tényezőktől mentesnek tűnnek, pedig valójában az adatok és az algoritmusok mögötti tervezés során beépített előítéletek, mint például a kisebbségekkel vagy különböző társadalmi csoportokkal szembeni diszkrimináció, komoly hatással lehetnek az eredményekre.

Az MI rendszerek alkalmazásai széleskörűek: használják őket orvosi diagnosztikában, pénzügyi elemzésekben, önvezető autók fejlesztésében, ügyfélszolgálati chatbotokban és számos más területen. Az üzleti életben, például egy hitelbírálat folyamatában, az ilyen típusú torzulások jelentős hátrányokhoz vezetik az érintett személyeket. Emellett az állásinterjúk során is előfordulhat, hogy a mesterséges intelligencia által végzett értékelés nem tükrözi pontosan az álláskereső képességeit és potenciálját, mivel az algoritmusok hajlamosak lehetnek az előítéletek reprodukálására.

A téma fontossága és korszerűsége miatt érdemes ezzel a témával foglalkozni és előtérbe helyezni a mesterséges intelligencia fejlesztése során, hiszen olyan döntéseket is hozhatnak a rendszerek, amik akár életeket is befolyásolhatnak, kezdve egy szabadlábra helyezéstől az önvezető autók morális dilemmájáig.

1.2. A dolgozat célja

A dolgozat célja, hogy a különböző elfogultság forrásai és kialakulásának helyei felismerésre kerüljenek. Az elfogultság felismerése mellet, szükségesek olyan módszereket is találni, ami felismeri és mérni is tudja ezen elfogultságok jellegét és mennyiségét, emellett olyan technikák kialakítása, ami csökkenteni tudja az elfogultság mértékét. Ennek eredménye hozzájárulhat egy olyan társadalom létrehozásához, ahol mindenki esélyt kap a tisztességes megítélésre és az igazságos lehetőségekre, függetlenül származásától, nemtől vagy egyéb társadalmi jellemzőtől.

Az ehhez tartozó kutatási terület szerintem kiemelkedően fontos, hiszen a technológiai innovációval nem kéne, hogy együtt járjon azzal, hogy egyes társadalmi rétegeket a XXI. században hátrányos megkülönböztetésben szenvedjenek. A mesterséges intelligenciai fejlődését a jogi és etikai jogszabályok nem tudja lekövetni, csak 5-10 éves csúszásban (Forum, 2020), ez azzal is jár, hogy a rendszerek alkalmazásakor a fejlesztőkben és a rendszer tulajdonosokban is meg kell bízniuk a felhasználóknak. Pont emiatt fontos, hogy a felhasználók tisztában legyenek, hogy milyen hátrányos megkülönböztetések érhetik a rendszer használatakor és milyen jogi lépéseket tehetnek ellene. A dolgozat főbb kérdései az alábbiak:

- Mit jelent, hogy torzított a mesterséges intelligencia?
- Feltételezzük-e, hogy nem kell tökéletesnek lennie? Ha igen, elég csak annyi, hogy jobb mint az ember?
- Hogyan lehet mérni a torzítottságot?
- Hogyan lehet elkerülni a torzítottságot?
- Milyen következményekkel járhat a torzított mesterséges intelligencia
- Hogyan befolyásolja az önvezető autókat a torzítottság?

A kutatás során elsődleges célom, hogy ezen kérdésekre választ találjak, valamint, ha felmerülne a kutatás során egyéb kérdéskör, akkor a dolgozat végén megfogalmazzam és választ adjak rájuk.

2. Mesterséges intelligencia alapjai

2.1. A mesterséges intelligencia meghatározása

A mesterséges intelligenciának nincs egy univerzális definíciója, hiszen nem csak egy friss kutatási területről van szó, hanem a folyamatos fejlődés, a széles alkalmazási területek és az interdiszciplináris jellege miatt sem alakult ki egyelőre. A mesterséges intelligencia fogalom a genus proximum mint intelligencia és a differentia specifica mint pedig a mesterséges jelzőből áll:

Az intelligencia meghatározásával többen is próbálkoztak és több álláspont is van ezzel kapcsolatban, viszont ami releváns a dolgozat szempontjából az a következő: Az intelligencia az a kognitív képesség, amely lehetővé teszi, hogy az egyén képes legyen tanulni és alkalmazkodni az új helyzetekhez. Ez magába foglalja, hogy az egyén információkat értelmez, logikailag gondolkodik és az ebből következtetett ismereteket alkalmazza.

A mesterséges szó olyan dolgokat foglal magába, amelyek nem természetes eredetűek, hanem az ember által lettek tervezve és/vagy létrehozva.

John McCarthy hozzájárult a mesterséges intelligencia fogalom bevezetéséhez, amit ő az intelligens gépek létrehozásának tudományaként írt le. Manapság mégis úgy írható le, hogy a mesterséges intelligencia nem egy biológiai organizmus, de mégis képes hasonlóan viselkedni, mint egy természetes intelligenciával rendelkező élőlény, emellett képes állandó emberi beavatkozás nélkül reagálni a környezeti változásokra és képes autonóm döntések meghozatalára. Az Európai Bizottság a következőket nyilatkozta a mesterséges intelligenciáról:

A mesterséges intelligencia fogalmát illetően nincs egységesen elfogadott nemzetközi megállapodás. A technológia jellemző jegyeinek kiemelésével elmondható, hogy olyan technológiák együttese, amelyek adatokat kombinálnak algoritmusokkal és számításiteljesítménnyel. (Európai Bizottság, 2020)

2.2. Történeti áttekintés

A mesterséges intelligencia több évtizedes kutatás eredménye, ami úgy ahogy a definícióból úgy a történelmében sincsen egyértelmű álláspont. Valamikor 1940/1950-re tehető az első kutatások egyike, amik a mesterséges intelligencia alapjai lettek. Kiemelkedik a kutatások közül Alan Turing Turning-tesztje. A Turning-teszt azt vizsgálja, hogy egy gép képes-e olyan intelligens viselkedésre, amely megkülönböztethetetlen egy embertől, azaz képes-e olyan válaszokat adni, mint egy ember.

„A teszt abból áll, hogy a bíráló billentyűzet és monitor közvetítésével kérdéseket tesz fel a két tesztalanynak, akiket így se nem láthat, se nem hallhat. A két alany egyike valóban ember, míg a másik egy gép, és mindketten megpróbálják meggyőzni a kérdezőt arról, hogy ők gondolkodó emberek. Ha a kérdező öt perces faggatás után sem tudja egyértelműen megállapítani, hogy a két alany közül melyik a gép, akkor a gép sikerrel teljesítette a tesztet" (Turing, 1950)

1956-ban összehívták a Dartmouth Workshop-ot, ahol John McCarthy bevezette a mesterséges intelligencia mint fogalmat, majd 1958-ban definiálta az elsődleges mesterséges intelligencia programozási nyelvet a Lisp-et.

„Sem meglepni, sem sokkolni senkit nem célom – de a legegyszerűbben összefoglalva azt mondhatom, hogy a világban léteznek ma már gondolkodó, tanuló és kreatív gépek. E képességük rohamosan fog fejlődni, és – a közeljövőben – az általuk feldolgozott problémák köre összemérhető lesz azokkal a problémákkal, amelyekkel az emberi elme eddig megküzdött." (Simon, 1980)

1960-ban Herbert Simon és Allen Newell megalkották a General Problem Solver másnéven a GPS-t. Ezt a programot alapból úgy tervezték, hogy imitálja az emberi problémamegoldás protokollját.

1970-ben "AI winter" időszak alakult ki a korlátozott technológiai előre lepések miatt emiatt csökkent a mesterséges intelligencia iránti érdeklődés egyúttal a finanszírozás is. 1970-1980-ban éledt újjá a neurális hálók iránti érdeklődés, ami az emberi agy működését próbálja lemásolni emellett ekkor fejlesztették ki a gépi tanulás során használt algoritmusokat például a döntési fákat, támogató vektor gépeket vagy a K-közép algoritmust. 1975-ben Patrick Henry Winston publikálta a "Learning Structural Descriptions from Examples" című munkáját, amelyben bebizonyítja, hogy a gépek képesek strukturált leírásokat tanulni példákból. (BME, a)

A sakk játszmák komplikáltsága és a több millió lépés kombinációjának köszönhetően több olyan modellt is fejlesztettek, ami a lehető legjobb ellenfele lehet az embernek. 1986-ban a Berliner HiTech sakkprogram elnyerte a nagymesteri címet majd 1988-ban Ed Formane világbajnokot legyőzte a sakkszámítógép.(BME, b)

A 2000-es évek óta megannyi fejlesztés és újítás jelent meg a piacon kezdve az önvezető autóktól át a hangvezérlésű rendszerekig például a Siri-ig vagy akár a 2023-as év legnépszerűbb termékéig a ChatGPT-ig.

A következő évtizedek és évszázadok olyan újításokat hoznak, amik egyelőre csak a képzeletekben léteznek, de a gyors technológiai fejlődésnek köszönhetően a közeljövőben megvalósíthatóak lesznek.

2.3. Mesterséges intelligencia és részterületei

A **mesterséges intelligenciával** olyan rendszerek létrehozása a cél, amik képesek olyan feladatokat elvégezni, amikhez emberi intelligencia felhasználása szükséges. Ilyen feladatok lehetnek döntéshozatali problémák vagy képfelismerés.
A **gépi tanulás** a mesterséges intelligencián belül egy részterület, amelynek célja, hogy adatokból tanulva fejlődjön és képes legyen feladatok elvégzésére anélkül, hogy explicit módon rá lennének tanitva ezekre. A gépi tanulás különböző algoritmusokat használ például felügyelt- , felügyelet nélküli- és megerősitéses tanulás.
A mesterséges intelligencián belül egy speciális ág a **mélytanulás**, amely a neurális hálókra alapul. A neurális hálók az emberi agy idegsejtjeinek mintázatára jött létre. Ezek a rendszer főleg komplex feladatok és problémák elvégzésére képesek, valamint hatalmas mennyiségű adatokban is felismerni mintázatokat és tanulnak ezekből.

2.4. A mesterséges intelligencia működési alapelvei

2.4.1. Tanulási módszerek

A mesterséges intelligencia különböző tanulási módszerekkel működnek, amik lehetővé teszik számukra az adatokból történő tanulást, döntéshozatalt és tapasztalat szerzést.

A **felügyelt tanulás** során a modell megkapja a bemeneti és a lehetséges kimenetei adatokat, vagyis az osztály címkéket. A címkék mellett a modellbe betáplálódik a tapasztalatok halmaza másnéven a tanító adatbázis. A cél, hogy a tanító adatokon betanult modell jól működjön az ismeretlen teszt adatokon is tehát megmondja az ismeretlen példához tartozó osztály címkéjét is. A mintákat jellemzőkkel vannak leírva és ezeknek a különbözősége és az osztály címkék korrelációjából következtethető egy még ismeretlen minta osztályára. Ezen jellemzők lehetnek véges diszkrét, valós vagy bináris értékűek.

Kétosztályos tanulás esetén minimális adat alapján is képes tanulni és következtetéseket levonni. A modell megtanulja azokat az általános mintákat és jellemzőket, ami segít az új feladat megoldásában. Az arcfelismerés során is használt tanulási modell, ahol a két bemenet beágyazóvektor dönti el, hogy hasonló vagy különböző-e a meglévő és a beovasott arc.

Többosztályos tanulás esetén N > 2 véges számú osztály címke van. A modellnek pedig a különböző jellemzők alapján kell besorolnia az ismeretlen adatokat a minta adatok alapján az egyes osztályokba.

A felügyelt tanulás kategóriájába tartozik a regressziós modell, amely segítségével egy vagy több független változó (inputok) és egy folytonos függő változó (output) közötti kapcsolat értelmezhető. A regresszión belül lineáris, többváltozós lineáris, polinomiális, logisztikus és sok más típus is létezik.

Lineáris regresszió esetén feltételezzük, hogy a függő és a független változók között lineáris kapcsolat van, ami az alábbi képlettel írható le:

$$y = b_0 + b_1 * x$$

ahol:

- az y a függő
- x a független változó
- b_0 az y -tengelyen lévő metszéspont
- b_1 pedig a meredekség

Többváltozós lineáris regresszió esetén több független változó van.

$$y = b_0 + b_1 x_1 + b_2 x_2 \pm \cdots + b_n x_n$$

A polinomiális regressziós esetén a kapcsolatot egy polinom függvény írja le.

$$y = b_0 + b_1 x + b_2 x^2 \pm \cdots + b_n x^n$$

Logisztikus regressziót a bináris osztályozási probléma megoldására alkalmas. A modell egy függő változó valószínűséget becsüli meg két osztályba sorolva.

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 \pm \cdots + b_n x_n)}}$$

A **felügyelet nélküli** modell esetében csak a bemeneti adatokat és a jellemzőket kapja meg, de a modellnek kell különböző logikák alapján csoportosítani és ismeretlen osztályokba sorolni a mintákat. Ez a módszer különösen hasznos, ha nincs elegendő előre definiált címkézett adat, vagy ha az adatok összetett mintázatainak és szerkezeteinek feltárása a cél.

A felügyelet nélküli tanuláson belül található a klaszterezési módszertan, aminek célja, hogy a hasonló egyedek egy csoportba, vagyis klaszterbe kerüljenek, amíg az ezekhez képest különböző egyedek más klaszterekbe. Klaszterezési algoritmusnak tekinthetjük többek között a K-közép, hierarchikus klaszterezést és DBSCAN-t is. A következőkben ezek lesznek kifejtve.

A K-közép (K-means) klaszterezési algoritmus célja, hogy megtalálja a klaszterek középpontját (centroidjait) és minden adatpontot a legközelebbi középponthoz rendelje. Az algoritmus működése során az adatok csoportosítására törekszik úgy, hogy a klaszterek középpontjai és az adatpontok közötti távolság minimális legyen. A centroidok és az adatpontok közötti távolságot az Eukleidészi távolsággal méri a modell, ez a későbbiekben lesz kifejtve.

Ezek után a modell újra inicializálja a centroidokat az adott klaszter összes adatpontjának átlaga kiszámításával:

$$C_i = \frac{1}{|N_i|} \Sigma \, x_i$$

Ezeket a lépéseket a modell addig ismétli, amíg a centroidok helyzete és az adatpontok becsült klaszterei nem változnak.

A K-közép algoritmus egy tovább fejlesztett változata a K-közép++. A K-közép++ az első centroidot véletlenszerűen választja ki majd a többi súlypontot már a maximális négyzetes távolság alapján az alábbi képlettel:

$$D_i = \max_{(j:1 \to k)} \|x_i - c_j\|^2$$

A k-közép++ inicializálás garantálja, hogy az algoritmus olyan megoldást talál, amely $O(\log k)$ versenyképes az optimális k-közép megoldással.

Az FCM algoritmus nem éles határokkal választja szét az adatpontokat, hanem minden adatpontot egy bizonyos fokú tagsággal rendel egy vagy több klaszterhez. Az algoritmus a centroidok véletlenszerűen választja ki majd egy mátrix létrehozásával feljegyzi, hogy egyes adatpontok, milyen mértékben tartoznak egy-egy centroidhoz. Ezek a tagsági értékek 0 és 1 közötti számok, amelyek valószínűségként értelmezhetőek. A klaszterek tagsági foka az alábbi képlettel

számolódnak ki:

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^m \cdot x_i}{\sum_{i=1}^{n} u_{ij}^m}$$

ahol:

- x_i az i-edik adatpont,
- u_{ij} az i-edik adatpont j-edik klaszterhez való tagsági foka,
- m pedig a fuzzy paraméter, ami általában 2 és szabályozza a fuzzy tagság mértékét.

Az algoritmus ezek után új centroid kiválasztása után újra kalkulálja a tagsági értékeket:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

ahol:

- $\|x_i - c_j\|$ az i-edik adatpont és a j-edik klaszterközéppont közötti távolság.

Amennyiben a tagsági fokok kisebb mértékben változtak, mint egy előre meghatározott érték abban az esetben megáll a kiválasztott centroidoknál, ha nem akkor pedig új centroidok lesznek.

Az algoritmus beállításához szükséges egy hiperparaméter, amely a klaszterek számát határozza meg, és ez a paraméter a modell tanítása során állandó marad. A klaszterek száma jelentősen befolyásolhatja az eredményeket, ezért fontos, hogy az ideális érték legyenek kiválasztva a modellezés előtt.

Hierarchikus klaszterezésnek két fő típusa van az agglomeratív és a divizív. Az agglomeratív hierarchikus klaszterezés során minden adatpont egy különálló klaszterként indul, majd a legközelebbi klaszterpárokat iteratívan összevonja a modell amíg csak egy klaszter marad. Az eredményeket általában egy dendogram nevezetű fa szerkezetben ábrázolják, amely bemutatja az összevonások sorrendjét és az egyes lépésekben mért távolságokat. A távolság mérésére három gyakori módszer áll rendelkezésre az Euklideszi, Manhattan és Koszinusz távolság. Az Euklideszi távolság a Minkowski távolságfüggvény leszármazottja.

$p = (p_1, p_2, \ldots, p_n)$ és $q = (q_1, q_2, \ldots, q_n)$ n-dimenziós térben

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

A Manhattan távolság vagy másnéven abszolút eltérés során a távolság abszolút értékét kell venni.

$p = (p_1, p_2, \ldots, p_n)$ és $q = (q_1, q_2, \ldots, q_n)$ n-dimenziós térben

$$d(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$

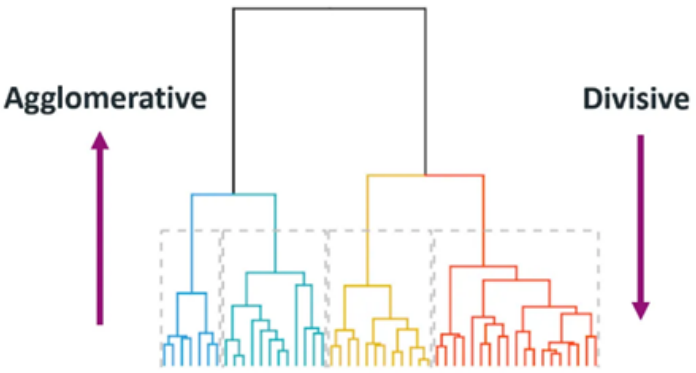
A Koszinusz hasonlóság két vektor közötti bezárt szög koszinuszát méri:

$$S_c = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

A koszinusz távolság pedig az 1- koszinusz hasonlóság. Az a és b vektor között a koszinusz hasonlóság a következőképp definiálható:

$$d_{\cos} = 1 - s_{\cos}(a,b)$$

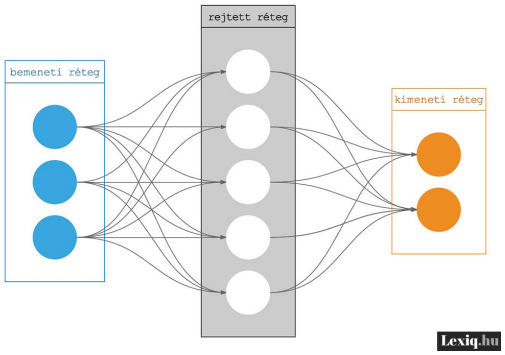
A diviszív hierarchikus klaszterben az összes adatpont egy klaszterből indul, amit iteratív módon kisebb klaszterekbe osztódik fel amíg minden adatpont nem kerül különálló klaszterekbe. Szintén dendogram diagrammban lehet ábrázol, de itt a felosztás sorrendjét mutatja be.



1. ábra. Hierarchikus klaszterezés
(Hadke et al., 2021)

A **neurális háló** algoritmus a biológiai agy működését utánozva tanul és hoz döntéseket. A modell több rétegben elrendezett mesterséges neuronból áll. amik hasonlóan működnek, mint az emberi agy idegsejtjei. Három jól elkülöníthető

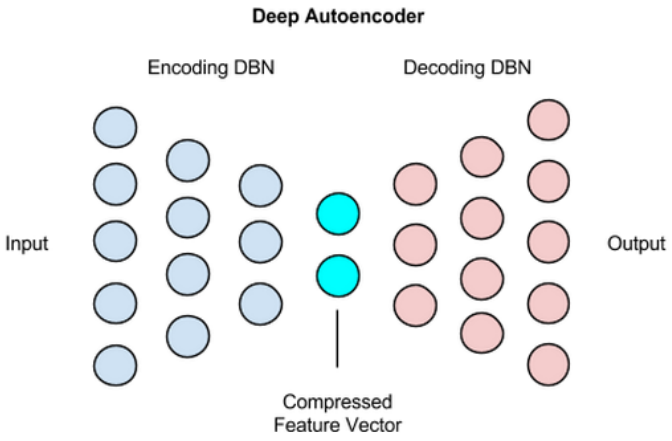
részből áll: bemeneti réteg(ek), rejtett réteg(ek), kimeneti réteg(ek).



2. ábra. Neurális háló felépítési ábrája
(Lexiq.hu, 2024)

A rejtett rétegek feladata, hogy transzformálják a bemeneti adatokat kimeneti adatokká a megfelelő logika szerint. A rétegek száma rendszerenként változhat és akár több száz réteg is kialakulhat, emiatt sem látható át ez a rendszerfolyamat és nem tekinthető teljes bizonyossággal optimálisnak a modell által létrehozott eredmény.

Az **autoencoder** a neurális háló egyik tipusa, ami elsősorban az adatok tömörítésére és zajmentesitésére használnak. Felügyelet nélküli tanulási folyamat során képes felismerni rejtett mintázatokat az adatok között anélkül, hogy címkéket alkalmazna. Az autoencoder két fő részből áll: az encoderből és a decoderből. Az encoder feladata, hogy az adatokat kisebb dimenziós térbe vetítse át, így egy sűrített reprezentációt hoz létre az eredeti adatokból. Az adat a háló egyre kisebb dimenziós rétegein halad keresztül, míg végül egy alacsonyabb dimenziós reprezentáció jön létre. A decoder feladata, hogy ezt a sűrített kódot felhasználva rekonstruálja az eredeti adatot, visszaállítva annak dimenzióit, hogy a lehető legpontosabb rekonstrukciót adja az eredeti adathoz képest.



3. ábra. Autoencoder működése
(Wiki, nd)

A PCA vagyis **főkomponens-analízis**, egy dimenziócsökkentési módszer, aminek célja, hogy nagy számú változót tartalmazó adahalmazt kisebb számú, de információs jellegét nem elveszitve új változókká vagyis főkomponensekké alakitsa át. Működésének lényege, hogy normalizálás után a kovariancia vagy korreláció kiszámításával feltárja a változók közötti összefüggéseket. A kovarianciamátrix elemzésével a sajátvektorok meghatározzák a főkomponensek irányát, míg a sajátértékek a főkomponensek által hordozott varianciát jelölik. Az első főkomponens az eredeti változók olyan lineáris kombinációja, amely a legnagyobb varianciát tartalmazza az adathalmazban. A második főkomponens szintén egy lineáris kombináció, amely a második legnagyobb varianciát hordozza, de az első főkomponenstől független (ortogonális). Ez a folyamat folytatódik, amíg meg nem lesz meg az összes főkomponenst. Miután a főkomponenseket meg lettek határozva, az adathalmaz kisebb dimenziós térbe kerül, ahol kevesebb főkomponens kerül megtartásra, mint ahány eredeti változó volt, de a főkomponensek továbbra is megőrzik az adathalmaz legfontosabb információit.

3. Az elfogultság típusai a mesterséges intelligenciában

A mesterséges intelligenciában gyakran előfordul, hogy egyes társadalmi rétegek vagy csoportokat diszkriminatívan érintenek egyes eredmény. Éppen ezért fontos, hogy a meghozott döntések jogszerűek, tisztességesek, átláthatóak és konzisztensek legyenek. Egyes lépések során el kell kerülni az olyan hibákat, amik befolyásolhatják az eredmények elfogultságát.

3.1. Miért alakul ki az elfogultság az emberekben?

3.1.1. Evolúciós előnyök

Az evolúció során a korai emberek túlélése érdekében alapvető fontosságú volt bizonyos minták gyors felismerése. Az ilyen minták közé tartozott például a ragadozók megjelenése, amely azonnali és hatékony reakciót igényelt a túlélés érdekében. Az agy fejlődése során kialakult az a képesség, hogy a környezetben megjelenő mintákat, különösen a potenciálisan veszélyeseket, gyorsan észlelje és értelmezze. Azok az egyének, akik képesek voltak a veszélyeket azonosítani és időben reagálni, nagyobb valószínűséggel élték túl, és így génjeik nagyobb valószínűséggel öröklődtek tovább az utódok számára.

Ez az evolúciós nyomás hozzájárult ahhoz, hogy az agyunk kifejlessze a mintafelismerés gyorsítására és azonnali reakciókra összpontosító mechanizmusokat. Azonban ez a mintafelismerési képesség nemcsak a veszélyek gyors észlelésére terjedt ki, hanem egyéb, a túlélés szempontjából releváns mintákra is, például a társas kapcsolatok és a csoportdinamikák kezelésére.

A diszkrimináció, vagyis a különböző ingerek észlelésének és megkülönböztetésének képessége, az agy fejlődésének egy másik aspektusát tükrözi. Az evolúciós nyomás arra ösztönözte az embereket, hogy gyorsan megkülönböztessék egymástól a potenciálisan barátságos és ellenséges csoportokat, amely segítette a társas csoportok hatékonyabb működését és a társas kapcsolatok optimalizálását. Azok, akik képesek voltak pontosan azonosítani a csoportok közötti eltéréseket, nagyobb eséllyel tudtak sikeresen navigálni a társas hierarchiákban és elkerülni a konfliktusokat.

3.1.2. Agy működése

Az evolúciós nyomás következtében az agy fejlődése során a mintafelismerési és reakcióképességi képességek gyorsabbá és hatékonyabbá váltak. Ezek a képességek különösen a temporális lebenyben és az amygdalában fejlődtek ki, amelyek a veszélyek és érzelmi ingerek gyors feldolgozásáért felelősek.

A mintafelismerés és a reakcióképesség gyakran automatikusan működött, amit a harc vagy menekülés reflex váltott ki. Ez a reflex a limbikus rendszerből származik, amely gyorsan aktiválódik veszélyhelyzetekben, elősegítve a gyors döntéshozatalt és cselekvést. Az agy ezen területei nemcsak a veszélyek gyors azonosításáért felelősek, hanem a társas interakciók során is kulcsszerepet játszanak, például a csoportok közötti diszkrimináció és a társas kapcsolatok kezelésében. Az agy ezen képességei lehetővé tették, hogy a korai emberek sikeresen navigáljanak a társas csoportok dinamikájában, elkerülve a potenciális konfliktusokat és kihasználva a társas előnyöket.

Az agy gyors mintafelismerő képességei és diszkriminációs képességei tehát nemcsak a közvetlen fizikai veszélyek kezelésére voltak fontosak, hanem a társas kapcsolatok és a csoportdinamikák hatékony kezelésében is kulcsszerepet játszottak. (LeDoux, 1998)

3.1.3. Kognitív torzítások

A kognitív torzítások, mint például a megerősítési torzítás és az első benyomás torzítás, az agyunk evolúciós örökségének fontos részei. Ezek a torzítások az emberi agy gyors és hatékony működésére utalnak, amely a túlélési szempontokkal kapcsolatos kihívások kezelésére fejlődött ki.

A **megerősítési torzítás** akkor jelentkezik, amikor hajlamosak vagyunk olyan információkat keresni és értelmezni, amelyek megerősítik meglévő hiedelmeinket, miközben figyelmen kívül hagyjuk azokat, amelyek ellentmondanak nekik. Evolúciós szempontból ez a torzítás segíthetett a gyors döntéshozatalban, mivel a korai emberek számára gyakran életbevágó volt, hogy gyorsan értékeljék a környezetükben megjelenő mintákat és információkat. Azok, akik hatékonyan és gyorsan integrálták a megerősítő információkat a meglévő tudásukba, jobban alkalmazkodtak a környezeti változásokhoz.

Az **első benyomás torzítás** akkor jelentkezik, amikor az első benyomások túlzottan befolyásolják a későbbi ítéleteinket és döntéseinket. Az evolúció során ez a torzítás előnyös lehetett, mivel a gyors első benyomás alapján történő döntéshozatal gyakran gyors reakciókat és alkalmazkodást igényelt. A korai emberek esetében az első benyomások gyors értékelése segíthetett a potenciális veszélyek, például egy ellenséges csoport vagy ragadozó gyors azonosításában.

3.1.4. Szocializáció és tanulás

A szocializáció és tanulás során az agy különböző területei aktívan részt vesznek a csoportidentitás kialakításában és fenntartásában.

Az emberi közösségekben a **csoportidentitás** kialakulása és a „mi" és „ők" közötti megkülönböztetés szintén evolúciós előnyöket jelentett. Az evolúciós előnyök között szerepelt, hogy az emberek képesek voltak kialakítani és fenntartani a csoportidentitást, ami hozzájárult a csoporton belüli együttműködéshez és szolidaritáshoz. Ez a kooperáció javította a csoport túlélési esélyeit, mivel a közös célokért való együttműködés és a közös védelem előnyös volt a csoport tagjai számára.

3.2. Mi a diszkrimináció?

A diszkrimináció szó jelentése a latin discrimino szóból ered, aminek jelentése megkülönböztetést, szétválasztás. Diszkriminációról van szó mikor egyes személyek megítélése nem a tetteik és jellemük alapján történik, hanem a társadalmi csoportjukhoz való tartozás alapján. A diszkrimináció nem csak negatív hatású lehet, hanem akár pozitív diszkriminációban is részesülhetnek az egyének vagy közösségek.

A diszkriminációnak két nagy fajtája van a megmagyarázható és a megmagyarázhatatlan. Az előbbi esetében statisztikailag megmagyarázható és igazolható az eredmény bizonyos tulajdonságok révén, még a megmagyarázhatatlan esetében a diszkrimináció indokolatlan, akár jogszerűtlennek is tekinthető.

A diszkrimináció forrása eredhet statisztikai eredményekből, mikor egyes csoportok statisztikai eredményét vetítik az egyénekre és ez alapján vonják le a következtetéseket. Valamint eredhet rendszerszintről is mikor szokásokra, politikai ideológiákra vagy viselkedési mintákra vonatkozik, amelyek egy szervezet kultúrájában mélyen gyökerezik és emiatt a társadalmi csoportok hátrányos helyzetbe kerülnek.

3.3. Elfogultság forrásai

Az elfogultság eredhet a modell által használt adatokból, valamint az ezekből levont téves következtetésekből. Fontos figyelni az algoritmusokra a programozás során, mivel már a tervezés fázisában is kialakulhat torzítás az adatokban.

3.3.1. Adatokból származtatható

Az adatokból származó elfogultság akkor lép fel, amikor a tanító adathalmaz torzítja a modell végeredményét, csökkentve ezzel a modell pontosságát és megbízhatóságát. Ezek a torzítások különböző forrásokból eredhetnek. Előfordulhat például, hogy a mérés vagy az adatszolgáltatás során az értékeket nem egységesen vagy nem megfelelően rögzítik a különböző csoportok esetében, ami torzítást okozhat. Ha a mintavételezés nem véletlenszerűen történik, az alcsoportok nem lesznek pontosan reprezentálva, ami hiányos adatokat eredményezhet. A reprezentációs hiba akkor jelentkezik, ha fontos változók kimaradnak az adatgyűjtésből, így az eredmények hiányosak és pontatlanok lesznek. Amennyiben a modell nem veszi figyelembe az egyéni különbségeket, és minden egyénre ugyanazokat a szabályokat alkalmazza, aggregációs torzítás léphet fel. Ezen kívül, ha a felhasználói tevékenységekből származó attribútumok nem tükrözik pontosan a felhasználók valódi viselkedését, kapcsolódási torzítás alakulhat ki. E torzítások különböző mértékben befolyásolhatják a modellek teljesítményét és megbízhatóságát, és fontos, hogy az adatgyűjtés és -feldolgozás során figyelembe vegyük őket a pontos és igazságos eredmények elérése érdekében.

Az adatmérgezés olyan mesterséges intelligencia ellen irányuló tevékenység amikor rosszindulatú támadók szándékosan hamis vagy káros adatokat injektálnak a rendszer által használt tanitó adatok közé. Az adatmérgezés történhet szándékosan is rosszindulatból, hogy ezzel befolyásolja a modell eredményeit és hátrányos helyzetbe kerüljenek bizonyos csoportok, valamint figyelmetlenség során is kialakulhat, mikor nem megfelelő és reprezentatív adatok kerülnek az adathalmazba.

Az adatméregzés remek példája, hogy a Google Térkép a telefonok helymeghatározását használva jelzi a forgalmat és a forgalmi dugókat. Simone Weckert ezt kihasználva 2020-ban 99 telefont és egy húzós talicskát használva okozott forgalmi dugót a Google Térkép szerint, miközben valójában dugóról szó sem volt. A telefonok helyadatait felhasználva a Google Térkép úgy érzékelte, hogy jelentős forgalmi dugó alakult ki az adott útvonalon, ez arra ösztönözte a járművezetőket, hogy kerüljék el ezt az útvonalat. Ez egyértelműen megmutatta, hogy hogyan lehet hamis adatokat bejuttatni a rendszerbe, amelyek félrevezető információkat eredményeznek. (Weckert, 2020)

3.3.2. Felhasználókból származtatható

Az adatokat sokszor a felhasználók generálják és abban az esetben, ha a felhasználók nem reprezentatív vagy valós adatokat szolgáltatnak abban az esetben a modell sem fog jól teljesíteni és reális eredményeket szolgáltatni.

A historikus torzítás a már meglévő társadalmi problémákból és konfliktusokból ered, ezek szolgáltatják a nem adekvát és nem reprezentatív adatokat a csoportokról és ezzel rontva a megítéléseket. Ugyanúgy fals statisztikai adatokat eredményezhet a modell amennyiben a célcsoportnak nem megfelelő, valamint akár önkéntes alapon történt az adatgyűjtés, így a valóságot nem reprezentálja megfelelően. A társadalmi torzítás mások cselekedeteinek hatására alakul ki, míg a viselkedési torzítás különböző platformokon vagy adathalmazokban eltérő felhasználói viselkedésből ered.

3.3.3. Algoritmusból származtatható

Algoritmikus torzításról akkor beszélünk, ha a torzítás nincs jelen a bemeneti adatokban, hanem az algoritmus maga adja hozzá. Ez történhet például az algoritmus tervezési döntései, optimalizálási függvények, regularizációk vagy a regressziós modellek hibás alkalmazása miatt.

A felhasználói interakciós torzítás akkor alakul ki, amikor a felhasználók saját viselkedésükkel torzítják az eredményeket. A népszerűségi torzítás pedig azt eredményezi, hogy a népszerűbb elemek gyakrabban jelennek meg a keresési eredményekben vagy ajánlórendszerekben, ami torzíthatja az eredményeket, különösen akkor, ha a népszerűségi mérőszámokat manipulálják.

Ilyen példa lehet a Netflix vagy az Amazon, amik, olyan ajánlórendszereket használnak, amelyek torzításokat hozhatnak létre. Például a felhasználóknak olyan tartalmakat ajánlhatnak, amelyek hasonlóak az általuk korábban látottakhoz, így megerősítve a már meglévő preferenciáikat és nem adva lehetőséget új, változatos tartalmak felfedezésére.

3.4. Elfogultság felismerése és kezelése

3.4.1. Elfogultság felismerése

Az elfogultság forrásai ismertek, de az, hogy az adat, az algoritmus vagy a modell valóban elfogult lesz-e, nem mindig egyértelmű. Ehhez különböző statisztikai mutatókat kell alkalmazni a kimeneti adatok minőségének felmérésére. Bináris modellek esetében az AUC mutató egy hasznos statisztikai mutató, de először meg kell ismerni a ROC görbét is.

A ROC görbe (Receiver Operating Characteristic curve) a bináris osztályozási modellek teljesítményének grafikus ábrázolása. Ez a görbe a valós pozitív arányt (TPR) ábrázolja a hamis pozitív aránnyal (FPR) szemben különböző küszöbértékek mellett

A TPR-t érzékenységi mutató vagy recall-ként is emlegetik. A mutató modell által helyesen pozitívnak prediktált érték aránya az összes pozitív arányhoz képest.

$$TPR = \frac{TP}{TP + FN}$$

FRP pedig azon negatív példák aránya, amelyeket tévesen pozitívnak azonosított a modell az összes negatív példához képest.

$$FPR = \frac{FP}{FP + TN}$$

Az AUC egy numerikus mérőszám, amely a ROC görbe alatti területet jelzi, ami azt adja meg, hogy alapvetően semleges szabályok mennyiben képesek adott csoportokat negatívan érinteni. Az AUC értéke 0 és 1 közötti intervallumon helyezkedhet el. Amennyiben az AUC = 1, abban az esetben a modell tökéletesen osztályoz, ha az AUC= 0.5, akkor a modell véletlenszerűen hoz döntéseket.

Másik statisztikai mérőszám lehet a DI (Disparate Impact), ami a védett csoportok és a referencia csoportok közötti arányt összehasonlításával számítják ki az alábbi képlet alapján:

$$DI = \frac{\text{Védett csoportok pozitív eredményeinek aránya}}{\text{Referencia csoportok pozitív eredményeinek aránya}}$$

A DI egy 0 és 1 közötti eredményt ad vissza, ahol az 1 a nincs aránytalan hatás, azaz a védett és referencia csoportok közötti esélyek egyenlőek. Amennyiben 1 > DI > 0.8 között van az értéke az Egyesült Államok Egyenlő Foglalkoztatási Bizottság (EEOC) döntésének alapján elfogadható a modell működése és amennyiben 0.8 alatt van abban az esetben jelentős az aránytalanság és ilyenkor a modellt szükséges módosítani.

Az EEOC egy további szabály alkalmazását javasolja, az úgynevezett 80%-

os vagy négyötöd szabályt. A szabály lényege, hogy a tanító adatban a védett csoportok aránya nem lehet a kiválasztási aránynál 80%-kal kisebb. Vagyis, ha egy csoport kiválasztási aránya nagymértékben alacsonyabb, mint a legmagasabb kiválasztási arány, akkor az eljárás aránytalan hatással lehet a védett csoportokra, és potenciálisan diszkriminatívnak tekinthető.

Az EEOC szerint a négyötöd szabály egyszerű és gyakorlati mérőszámot nyújt a diszkrimináció felismeréséhez. Az ilyen aránytalanságokat azonosítani kell, hogy a szervezetek biztosítani tudják a tisztességes és méltányos eljárásokat, valamint az esélyegyenlőséget minden csoport számára. Amennyiben a DI mutató értéke 0,8 alatt van, a gyakorlatban gyakran szükség van további vizsgálatra és a modell módosítására, hogy csökkentsék az aránytalan hatást és növeljék a méltányosságot a döntéshozatali folyamatban.

Ezen túlmenően, ha a 80%-os szabály nem teljesül, az érintett szervezeteknek ki kell vizsgálniuk a folyamatokat és a döntéshozatali kritériumokat, hogy megértsék a diszkrimináció okait, és szükség esetén módosítaniuk kell azokat a tisztességesebb eredmények elérése érdekében.

3.4.2. Elfogultság kezelése

Az adatgyűjtés és -feldolgozás során kritikus kérdés, hogy milyen csoportokat definiálunk és milyen mélységig fúrunk le a társadalomban. Például, egy csoport meghatározása lehet etnikai alapon, mint a feketék, vagy lehet komplexebb, több szempontot figyelembe vevő, mint például afroamerikai zsidó emberek alacsony termettel halmaza. Az ilyen specifikus csoportok meghatározása segíthet abban, hogy a modell érzékeny legyen a különböző társadalmi rétegek sajátos igényeire és jellemzőire.

Ezen csoportok azonosítására és a diverzifikáció mértékének vizsgálatára információs kritériumokat, például entrópiát lehet alkalmazni. Az entrópia mérése segíthet annak megértésében, hogy mennyire változatosak az adatok egy adott csoportban, és hogy milyen mértékben vannak reprezentálva különböző szegmensek. Minél magasabb az entrópia, annál nagyobb a diverzitás az adott adathalmazban, ami hozzájárulhat a modell méltányosságához és általánosíthatóságához.

Az entrópia mellett más mérőszámokat és módszereket is használhatunk, mint például a Gini-koefficiens vagy a Shannon-index, amelyek segítenek az adatok sokszínűségének és egyensúlyának elemzésében. Ezek a módszerek lehetővé teszik, hogy az adatgyűjtési és -feldolgozási folyamatok során folyamatosan ellenőrizzük a reprezentativitást és minimalizáljuk az esetleges torzításokat.

Shannon-entrópia:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

ahol:

– $H(X)$ az entrópia,

– $p(x_i)$ a x_i esemény bekövetkezésének valószínűsége,

– n az összes lehetséges események száma.

Gini-koefficiens:

$$G = 1 - \sum_{i=1}^{n} (P(x_i))^2$$

ahol:

– G a Gini-index,

– $P(x_i)$ a x_i esemény bekövetkezésének valószínűsége,

– n az összes lehetséges események száma.

A **sziluett** érték a klaszterezési algoritmusok teljesítményének értékelésére használható. A klaszteren belüli szorosság és a többi egyedtől való távolságtartás mértéke.

Sziluett érték:

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}$$

ahol:

– a(x) az átlagos távolság x mintaegyed, és az összes vele egy klaszterben lévő egyed között.

– b(x) az átlagos távolság x mintaegyed és az összes, egyéb klaszterben lévő mintaegyed között.

A sziluett értéke:

– 1 -hez közeli, ha a pont jól illeszkedik a klaszterbe és távol van a többi klasztertől, azaz az algoritmus közelíthetően tökéletesen osztályoz.

– 0 körüli érték, ha a pont a klaszter határán helyezkedik el, azaz az algoritmus közömbösen osztályoz.

– -1 -hez közeli, ha a pont inkább egy másik klaszterbe illeszkedik, vagyis az algoritmus közelíthetően teljesen félreosztályoz.

Végső soron az adatgyűjtés és adatfeldolgozás során alkalmazott módszerek meghatározása és finomhangolása alapvető fontosságú a gépi tanulási modellek teljesítményének és méltányosságának biztosítása érdekében.

Az adatgyűjtés során fontos biztosítani a torzítottság mentességét. Ehhez elengedhetetlen, hogy az adatok diverzifikáltak legyenek, vagyis különböző forrásokból származó, sokszínű és reprezentatív adatbázist kell létrehozni. Szükséges a hibás, torzított, illetve duplikált adatok eltávolítása, valamint a csoportok közötti egyensúly megteremtése az adatfeldolgozás során. A diverzifikált és kiegyensúlyozott adatok minimalizálják az elfogultságot, és biztosítják, hogy a modell általánosítható legyen minden csoport számára. Az adatfeldolgozás technikái között szerepelhet a minták kiegyensúlyozása is, például az alul reprezentált csoportok adatainak növelése vagy a túlreprezentált csoportok adatainak csökkentése. Fontos, hogy az adatgyűjtés és feldolgozás módszerei alapvetően befolyásolják a modell későbbi teljesítményét és méltányosságát.

A tanítási folyamat előtt kritikus a megfelelő jellemzők (features) kiválasztása annak érdekében, hogy elkerüljük a torzított változók hozzáadását, amelyek növelhetik az elfogultságot, vagy éppen olyan változók kihagyását, amelyek segítenék a modell helyes működését.

Az algoritmusok kiválasztásakor kiemelten fontos olyan modelleket preferálni, amelyek figyelembe veszik az igazságosságot és minimalizálják az elfogultságot, az úgynevezett fairness-aware algoritmusokat. Az IBM AI Fairness 360 eszköze és a Google Fairness Indicators támogatást nyújtanak az elfogultság felismerésében és a modellek méltányosságának növelésében. Ezen kívül szükséges olyan technikák alkalmazása is, amelyek nehéz döntési helyzeteket hoznak létre szándékosan az algoritmusok számára, ezáltal tanítva azokat az elfogult esetek kezelésére.

Az algoritmusok átláthatósága nélkülözhetetlen nemcsak jogi szempontból, hanem azért is, hogy a fejlesztők megértsék a döntéshozatali mechanizmusokat és szükség esetén javítani tudják azokat.

Amennyiben a fejlesztők tudják, hogy milyen kimeneti adatokat kell visszakapniuk az egyes bemeneti adatokból, lehetőség van tesztek írására. Ezek lehetnek Unit tesztek, amelyek az egyes komponensek külön-külön történő tesztelését jelentik, biztosítva a részek optimális működését. Az integrációs tesztek pedig az

egyes modulok közötti integrált együttműködés biztosítására szolgálnak. A fairness tesztek kifejezetten az elfogultság és igazságosság vizsgálatát szolgálják, összehasonlítva különböző csoportok kimeneti eredményeit, hogy azonosítsák az esetleges aránytalanságokat. A validációs tesztelés során pedig ellenőrzik a modell kimeneti adatainak megfelelőségét, figyelembe véve a várt kimeneti adatokat. Ha ismertek a kimeneti adatok, lehetőség van utófeldolgozásra, ahol finom hangolják a modelleket.

Az elfogultság kezelésének egyik hatékony módja, hogy multidiszciplináris csapatok dolgoznak együtt. Így biztosítják, hogy az AI rendszerek ne csak technikailag, hanem társadalmilag is relevánsak és elfogadhatóak legyenek, bevonva például szociológusokat, etikusokat és más területek szakértőit is.

3.5. Jogi szabályozásai az Európai Unióban

3.5.1. Jogi kérdések

A mesterséges intelligencia jogi szabályozása az elmúlt pár évben kezdődött el és jelenlegi formájában nagyon kis részét fedi le a felhasználási köröknek. A felmerülő jogi kérdéseknek és a folyamatos fejlődés miatt nem is lehet lefedni az egész kutatási területet. Általános jogi elvárások a mesterséges intelligencia rendszerek felé, hogy legyenek függetlenek és pártatlanok, jogilag biztosak és következetesek és mindenkinek biztosítva legyen a törvény előtti egyenlőség. Az Európai Unióban több jogi szabályozás is törvénybe lépett, amik ezen elvárások betartására kényszerítik a fejlesztőket ezek mellett pedig az alapjogokat és az adatok védelmét is biztosítja.

A mesterséges intelligencia rendszerek több Európai Uniós Alapjogi Charta cikket is megszeghetnek, többek között a magán- és családi élet tiszteletben tartását (7.cikk), személyes adatok védelmét (8.cikk) , véleménynyilvánítás és a tájékozódás szabadságot (11.cikk), megkülönböztetés tilalmát (21.cikk) és a hatékony jogorvoslathoz és a tisztességes eljáráshoz való jogot (47.cikk).

Az első Európai Uniós mesterséges intelligenciára vonatkozó szabályozási javaslatot 2021 áprilisában tette az Európai Bizottság. Egy olyan keretrendszert alkalmazását szorgalmazták, amiben a mesterséges intelligenciát használó rendszereket különböző osztályokba sorolják, annak mérten, hogy a felhasználókra milyen mértékű kockázatot jelent. Az ”AI Act” néven elhíresült szabályozás végül 2024 tavaszán fogadta el az Európai Parlament és 2024. augusztus elsején lép életbe. Egy 24 hónapos átmeneti időszak után 2026 közepétől lesz az európai és a nemzeti jogrendszer része.

A jogszabály a világon az első átfogó szabályozás a mesterséges intelligencia területén. Az AI Act kategorizálja a kockázati szinteket mesterséges intelligencián belül: elfogadhatatlan, magas, alacsony és általános kockázat között.

Az első kategória az elfogadhatatlan kockázat, ami veszélyt jelent az emberekre és azonnali betiltásra kerülnek ezek a rendszerek vagy alkalmazások. Amennyiben a rendszer tartalmaz az emberekre és/vagy veszélyeztette csoportokra ható kognitív viselkedési manipulációt, ami akár veszélyes viselkedésre is motiválhat, abban az esetben azonnali betiltásra kerül. Az emberek osztályozása társadalmi vagy gazdasági státusz alapján is tiltott, valamint a (valós idejű) biometrikus azonosítás is.

Magas kockázatú rendszerek/alkalmazások közé tartoznak, amik a biztonságot vagy az alapvető jogokat negatívan befolyásolja, ezen rendszereket folyamatosan ellenőrzik és felülbírálják a besorolás szempontjából. Ezeket két kategóriába lehet sorolni a mesterséges intelligencia rendszereket, amiket az Európai Unió termékbiztonsági jogszabály hatálya alá tartoznak. A másik kategória nyílt speciális területet tartalmaz, amit regisztrálni kell az uniós adatbázisba, ezek a következők: kritikus infrastruktúra kezelése és üzemeltetése, oktatás, foglalkoztatás és munkavállalók irányítása, hozzáférés az alapvető magánszolgáltatásokhoz és a közszolgáltatásokhoz és azok előnyeihez, bűnüldözés, migráció és határellenőrzés, segítségnyújtás a jogértelmezésben és a jogalkalmazásban.

Ebbe a kategóriába tartozó rendszereknek szigorú követelményeknek kell megfelelniük:

- Részletes dokumentáció készítés
- Tevékenység naplózása
- Nagyfokú robusztusság, biztonság és pontosság
- Felhasználó egyértelmű és megfelelő tájékoztatása

Az általános célú rendszerek esetében a legnagyobb kockázatot az átláthatatlanság jelenti, ezért a jogszabály ezen megszüntetésére törekszik, bár ezek a rendszer nem minő magas kockázatúnak.

Amennyiben a rendszer az alacsony kockázatba lett besorolva, abban az esetben nincs semmilyen jogi kötelezettsége. A rendszer tulajdonosai és fejlesztőik önkéntesen dönthetnek arról, hogy alkalmazzák a jogszabályi elő írásokat. European Parlament (2024)

Az Európai Uniós Általános Adatvédelmi Rendelet vagy köznapibb nevén GDPR szintén szabályozza az automatizált döntési folyamatokat például, amik a gépi tanulásra épülnek. A profilalkotást az alábbiakban fogalmazzák meg:

„Személyes adatok automatizált kezelésének bármely olyan formája, amelynek során a személyes adatokat valamely természetes személyhez fűződő bizonyos személyes jellemzők értékelésére, különösen a munkahelyi teljesítményhez, gazdasági helyzethez, egészségi állapothoz, személyes preferenciákhoz, érdeklődéshez, megbízhatósághoz, viselkedéshez, tartózkodási helyhez vagy mozgáshoz kapcsolódó jellemzők elemzésére vagy előrejelzésére használják." (GDPR, 4.cikk, 4.pont)

Az általános tilalmat, ami a profilalkotást is tartalmazza az alábbi jogszabály biztosítja:

„(1) Az érintett jogosult arra, hogy ne terjedjen ki rá az olyan, kizárólag automatizált adatkezelésen – ideértve a profilalkotást is – alapuló döntés hatálya, amely rá nézve joghatással járna vagy őt hasonlóképpen jelentős mértékben érintené."(GDPR, 22.cikk, 1.bekezdés)

Ezen jogszabály alól kivételt képeznek azok az esetek, ahol érvényesülnek az alábbiak:

- *„a) az érintett és az adatkezelő közötti szerződés megkötése vagy teljesítése érdekében szükséges;*
- *b) meghozatalát az adatkezelőre alkalmazandó olyan uniós vagy tagállami jog teszi lehetővé, amely az érintett jogainak és szabadságainak, valamint jogos érdekeinek védelmét szolgáló megfelelő intézkedéseket is megállapít; vagy*
- *c) az érintett kifejezett hozzájárulásán alapul." (GDPR, 22.cikk, 2.bekezdés)*

A GDPR-ban az adatkezelők felelősségét és kötelességét is rögzítették, többek között a 71. preambulumbekezdésben. Itt az automatizált döntéshozatalban szükséges az ideális matematikai és statisztikai eljárásokat alkalmazni, megakadályozni a hátrányos megkülönböztetést, valamint olyan műszaki és szervezési intézkedéseket kell bevezetni, ami minimálisra csökkenti a hibalehetőséget és előidézi az adatok pontatlanságának korrekcióját.

3.5.2. Etikai kihívások

A felhasználók részéről is rengetek elvárás van a mesterséges intelligencia rendszerek felé, hogy a bizalom megfelelő és alátámasztott legyen. A rendszereknek megmagyarázhatónak, biztonságosnak, átláthatónak, számonkérhetőnek és tisztességesnek kell lennie. Az utóbbi három elvárás a legnehezebben teljesíthető. A tisztességességről és annak fontosságáról már esett szó korábban így a következőkben a másik kettőről lesz szó.

Az **átláthatóságot** nem csak a felhasználók igénylik, hanem a GDPR-ban is több jogszabály vonatkozik erről, hiszen nagyon fontos tudni azokat a folyamatokat és döntési elveket, amiket az algoritmusok hoznak, így ellenőrizve, hogy valóban nem történt torzítottság a döntések meghozatalakor. A jogszabályok betartása során és a felhasználók számára fontos, hogy a működési mechanizmus ezáltal az algoritmus is nyilvános legyen, így mindenki számára transzparenciát nyújtana, de felmerül a kérdés, hogy amennyiben ezek az algoritmusok nyilvánosak, hol éri meg ez a tervezőnek.

A nyilvánosságra hozatal ugyanis versenyelőny elvesztésével járhat, mivel a szellemi termékek védelme nélkül a konkurencia könnyebben lemásolhatja és felhasználhatja azokat. Ez pénzügyi veszteséget okozhat a tervezőknek, és növelheti a kibertámadások és visszaélések kockázatát is.

Fontos ugyanakkor megjegyezni, hogy nem minden algoritmus esetében jelent problémát a transzparencia biztosítása. Például a teljes Linux kernel forráskódja nyilvánosan elérhető a GitHubon, és ennek ellenére az egyik legbiztonságosabb szoftvernek számít, amely számos orvosi és ipari eszközön fut. A felhasználók jelezhetik az esetleges hibákat vagy veszély forrásokat a programozók felé, hiszen a felhasználónak is célja, hogy minnél biztonságosabb rendszert tudjon használni. Ez mutatja, hogy a nyilvánosság és a biztonság nem zárják ki egymást feltétlenül.

A feketedoboz effektus azonban valódi kihívást jelent, mivel számos algoritmus esetében nehéz vagy lehetetlen megérteni a belső működést és a döntéshozatali folyamatokat. Ez különösen igaz a gépi tanulás és a mesterséges intelligencia alapú rendszerekre, ahol az algoritmusok gyakran bonyolultak és nem átláthatóak.

A fentiek figyelembevételével javasolt egyensúlyt találni a transzparencia és a biztonság között, valamint olyan szabályozási kereteket kialakítani, amelyek biztosítják a felhasználók jogainak védelmét anélkül, hogy a tervezők versenyelőnyét és innovációs képességét jelentősen csökkentenék. Ebben a kontextusban

felmerülhet a kérdés, hogy az állam milyen szerepet vállalhatna a felülvizsgálatban és ellenőrzésben, de ez újabb kihívásokat és kérdéseket vet fel, amelyek alapos megfontolást igényelnek.

A **fekete doboz** jelenséget először 1941-ben Wilhelm Cauertől eredeztethető, aki kidolgozta az elméleti hátterét ennek a jelenségnek. A feketedoboz jelenség során a modell egy bemeneti adatot (input) kap és egy kimeneti adatot (output) ad vissza, de a folyamat és a döntési mechanizmus nem ismert, emiatt a modell produkálhat olyan eredményt, amelyre nincsen semmilyen magyarázat. A fekete dobozt bármire rá lehet húzni, legyen az tranzitor, algoritmus vagy emberi agy. A neurális hálózat során írható le a legjobban ezt a jelenséget.

4. ábra. Feketedoboz jelenség
(Saját szerkesztés)

Az átláthatatlanság miatt az adatkezelés sem tekinthető megfelelőnek, magyarul nem felel meg a GDPR-nak. Amennyiben sikerülne feltárni ezeket a rejtett rétegeket és bizonyítottan megfelelne az átláthatósági feltételeknek, abban az esetben használhatók lennének a neurális hálók, de addig a döntési fán alapuló algoritmusokra kell támaszkodnia a rendszereknek.

Az élet folyamán rengetegszer kell szembesülni, hogy a döntésekért felelősséget kell vállalni, ez a mesterséges intelligencia során, hogy ki a felelőse egyes döntésekért nem annyira egyértelmű. A **számonkérhetőség** ezért kiemelten fontos kérdés az MI rendszerek során. A hagyományos elszámolhatósági sztenderdeket és eljárásokat emberi döntéshozatalra alakították ki, emiatt sokszor nem is lehet kezelni az automatizált döntések által felvetett új kérdéseket. A felelősök köre is elég tág hiszen a fejlesztők, az üzemeltetők, rendszer tulajdonos vagy akár a felhasználók is beletartozhatnak. Mindenkinek meg van a saját felelősségi köré, de mégsem vonható mindenki felelősségre, hiszen a feketedoboz jelenség itt is megjelenik. Az átláthatatlan döntési mechanizmus miatt a felelősség is átláthatatlan, hiszen nem feltétlen ember által okozott hibákról van szó, hanem olyan rendszerhibáról, aminek eredete sem feltétlen mondható meg.

Új jogszabályok megalkotásával és a felelősség kijelölésével, sokkal jobban biztosítva lenne, hogy a rendszerek működése valóban megfeleljenek az elvárásoknak. Ez különösen fontos az alábbi területeken:

1. A mesterséges intelligencián alapuló rendszerek fejlesztésének korai szakaszában fontos, hogy szigorú tesztelési és validációs eljárásokat alkalmazzanak. A fejlesztőknek felelősséget kell vállalniuk az általuk létrehozott algoritmusok és modellek pontosságáért, megbízhatóságáért és etikai megfelelőségéért. A fejlesztési dokumentációk és a tesztelési eredmények nyilvántartása segíthet az átláthatóság növelésében.

2. Az üzemeltetők és rendszer tulajdonosok felelőssége, hogy a rendszereket folyamatosan monitorozzák és karbantartsák. Az anomáliák és hibák gyors felismerése és javítása kulcsfontosságú a rendszer megbízhatóságának fenntartásában. Az üzemeltetőknek biztosítaniuk kell, hogy a rendszerek megfeleljenek a biztonsági és adatvédelmi előírásoknak.

3. A felhasználók felelőssége, hogy megfelelően használják a rendszereket, és tisztában legyenek azok korlátaival. Fontos, hogy a felhasználók optimális képzést kapjanak a rendszerek használatáról, és tudják, hogyan reagáljanak a rendszer esetleges hibáira vagy nem várt viselkedésére.

4. Az új jogszabályoknak világosan meg kell határozniuk a különböző szereplők felelősségi körét. A jogi kereteknek lehetőséget kell biztosítaniuk arra, hogy a felelősöket számon lehessen kérni a rendszer hibái vagy helytelen működése esetén. Emellett a szabályozásoknak ösztönözniük kell a fejlesztők és üzemeltetők közötti együttműködést, valamint a legjobb gyakorlatok megosztását.

3.5.3. Ember és gép kapcsolata

Felmerül a kérdés, hogy miért jobb az, ha a mesterséges intelligencia végez el bizonyos döntéseket vagy számításokat az ember helyett.

A nyilvánvaló előnyök közé tartozik a mesterséges intelligenciánál, hogy sokkal nagyobb adatmennyiséget sokkal rövidebb idő alatt képes feldolgozni és ezáltal, olyan mintázatokat és következtetéseket tud, felfedni és levonni, ami kisebb adatmennyiségen, több erőforrást igényelve nem feltétlenül rajzolódna ki. A monoton feladattűrés is a mesterséges intelligencia javára írható, hiszen képes ugyanazt a feladatot újra és újra pontosan elvégezni anélkül, hogy elfáradna vagy hibázna. A gépi tanulás révén a mesterséges intelligencia képes a múltbeli adatokból tanulni és fejlődni, ezáltal egyre jobb teljesítményt nyújtani az idő múlásával.

Felmerül előnynek, hogy a mesterséges intelligenciát nem megzsarolható, megfenyegethető, megvesztegethető, nem korrupt és nem is szenved előítéletektől és ezekből adódóan kivételezni sem fog senkivel szemben sem. Ezen indokok mellett előnye, hogy ugyanabban az esetben ugyanazokat a döntéseket fogja meghozni, ez az ember esetében nem feltétlen elmondható. Az Egyesült Királyságban egy tanulmány során 81 bírót kérdeztek meg, hogy milyen óvadékot szabnának ki számos képzeletbeli vádlottnak. Minden képzeletbeli bűncselekménynek volt háttér története is. A 41 történetből hét történet kétszer is megjelent a bírók elött más névvel, hogy a bírók ne vegyék észre a duplikációkat. A legtöbb bíró esetében nem ugyanaz a döntés született másodjára, mint ami az első esetben (Fry, 2018). Ez a tanulmány is alátámasztja az állítást, hogy a gép sokkal konzisztensebb, mint az ember.

Az ember által hozott döntések, lehetnek diszkriminatívak, még akkor is, ha ez tudat alatt történik meg. A mesterséges intelligenciát ezek a tényezők nem befolyásolják, csak abban az esetben, ha ezek az érzelmek és előítéletek bele lettek táplálva a tanulás során.

A mesterséges intelligencia hátránya a felsorolt etikai kérdések mellett, hogy sokszor az internetről tanul, amin nincsenek szűrve az adatok ezért a hamis információk mellett, diszkriminatív adatokból is tanul, ami hatással lesz a döntésmeghozatalra is.

Az egyre felkapottabb ”Explainable AI” vagyis megmagyarázható mesterséges intelligencia célja, hogy olyan módszereket és technikákat alkalmazzon, amelynek a kimenetelei érthetőek és átláthatóak az emberek számára. Az Explainable AI példákat mutat a múltból, hogy milyen inputokra milyen outputok voltak az eredmények, vizualizálja a döntési lépéseket, természetesen nyelven elmagyarázza a döntések miértjét és elmagyarázza, hogy az egyes jellemzők, milyen mértékben és irányba befolyásolták a döntést. Ez az újítás nagymértékben befolyásolja majd a döntési mechanizmusokat.

4. Önvezető autók morális dilemmái

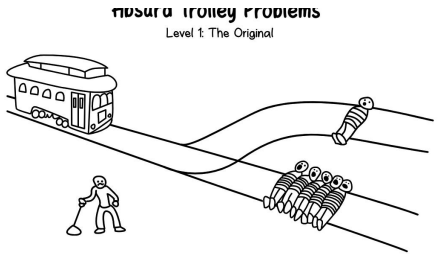
Az önvezető autók megjelenése forradalmi változásokat ígér a közlekedés terén, ígéretes megoldásokat kínálva a közlekedési balesetek számának csökkentésére, a közlekedési dugók enyhítésére és a mobilitás növelésére. Azonban e technológiai újításokkal együtt komoly etikai kérdések és morális dilemmák is felmerülnek, amelyek nemcsak a fejlesztők, hanem a társadalom egésze számára is kihívást jelentenek.

4.1. A morális dilemmák alapjai

Az önvezető autók esetében nagyon sok döntési helyzet van mikor nyilvánvalóvá válnak ezek a morális kérdések. Ezek a döntési helyzetek, potenciálisan veszélyt jelenthet minden olyan személyre nézve, aki részt vesz a közlekedésben. Ilyen döntési helyzet például, mikor az autó ütközés előtt áll és milyen prioritás szerint dönt, hogy kit vagy mit óvjon meg. Esetleg az autóban ülők élete fontosabb, mint a gyalogosoké, vagy pont fordítva? Erre a kérdésre alapozva, jött létre egy kutatás villamos-probléma (trolley problem) néven.

4.1.1. Villamos probléma

Phillippa Foot filozófus 1967-ben dolgozta, ki a villamos problémának nevezett etikai dilemmát. A dilemma kérdése, hogyan döntsünk, olyan helyzetekben, ahol nem lehet jó döntést hozni, csak csökkenteni lehet a károkat. A legnépszerűbb villamos probléma a következő képen van szemléltetve:



5. ábra. Villamos probléma
(Pancake, 2022)

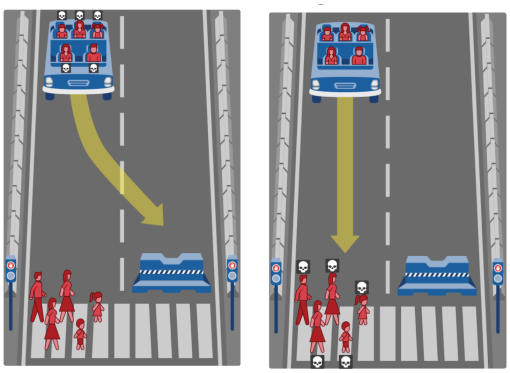
A probléma lényege, hogy egy irányíthatatlan villamos halad a síneken, és ha semmi nem változik, öt embert fog elütni. A válaszadónak van lehetősége átváltani a váltót, hogy a villamos egy másik vágányra fusson, ahol csak egy ember áll, akit így elütne a villamos. Egy 2001-es (Greene et al., 2001) kutatás szerint a válaszadók többsége utilitarista módon gondolkodott és hozta meg a döntését. Az utilista gondolkodás szerint az a helyes cselekedett, amely a legnagyobb boldogságot okozza a többi ember számára, ebben az esetben az öt ember életének a megmentése. Ha valaki valóban ilyen döntési helyzetbe kerülne, valóban így döntene vagy ez csak egy társadalmi kényszer, aminek meg akar felelni?

A fenti példát követve, ha nem a váltót kéne állítani, hanem egy embert a válaszadónak, saját kézzel kéne lelöknie a hídról, hogy megmentsen öt embert, akkor megtenné? Ebben az esetben a válaszadók, sokkal kisebb százaléka válaszolt igennel ugyanis ez már sokkal személyesebb, mint egy váltókart meghúzni, pedig mindkét esetben ugyanaz a végeredmény.

Egy kutatás szerint az emberek 76% gondolja, úgy, hogy minél több életet mentsenek meg az önvezető autók, még ha ez az autóban ülők életébe is kerülne, viszont mikor a résztvevőket megkérdezték, hogy vásárolnának-e olyan autót, ami adott körülményekben megölné őket, akkor már nem voltak hajlandóak feláldozni magukat a nagyobb jó érdekében. (Bonnefon et al., 2016)

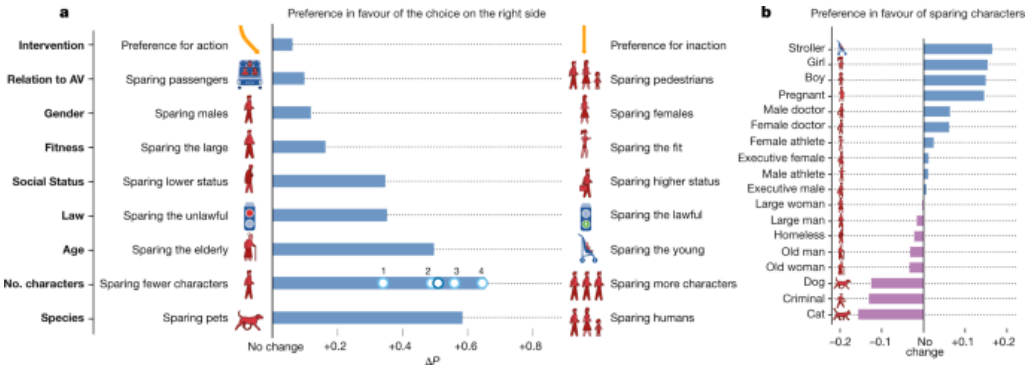
4.1.2. Moral Machine

Az MIT által létrehozott Moral Machine kutatás a villamos problémára épül és az önvezető autókra fókuszál. A kérdések során nem, olyan jellemzők is felmerülnek, hogy valaki a piros lámpán halad-e át, hogy fit vagy túlsúlyos-e a gyalogos, babakocsival van-e, fiatal vagy öreg, nő vagy férfi, állat vagy ember, mi a munkája, vagy pedig az autóban hányan ülnek.



6. ábra. Moral machine
(Moral Machine, 2024)

Egy 2018-as Nature kutatás (Awad et al., 2018) eredménye a következő lett:



7. ábra. Nature kutatás eredménye
(Gordon et al., 2018)

A kutatásban részt vettek szerint legnagyobb százalékban a babakocsival közlekedőket mentenék meg, előnyben részesítve a gyalogosokat az autóban utazók-

kal szemben, a nőket a férfiakkal szemben, a sportosabb testalkatúakat az idősebb vagy túlsúlyos emberekkel szemben, a magasabb státuszúakat a társadalmilag alacsonyabb státuszúakkal szemben, a fiatalokat az idősebbekkel szemben, több ember megmentését kevesebb emberrel szemben, az embereket az állatokkal szemben, és azokat, akik a közlekedési szabályokat betartva a zöld lámpánál közlekednek a szabályszegőkkel szemben.

Ez a kutatás felveti az etikai dilemmák és prioritások kérdését, mivel az emberek döntései különböző csoportokat részesítenek előnyben, amelyek megítélése etikai és igazságossági szempontból is kérdéseket vet fel. A kulturális különbségek is befolyásolhatják ezeket a döntéseket, mivel más országokban vagy kultúrákban eltérő eredmények születhetnek, ezek lentebb lesznek kifejtve. Az önvezető autók programozása során felmerül a kérdés, hogy hogyan kellene ezeket a járműveket úgy programozni, hogy etikusan cselekedjenek vészhelyzetekben, és hogy az emberek véleménye alapján kellene-e dönteni, vagy más etikai elveket kellene követni.

A társadalmi egyenlőség szempontjából az eredmények rámutatnak arra, hogy az emberek különböző társadalmi státuszú és egészségi állapotú személyeket másképp ítélnek meg, ami előítéleteket és diszkriminációt tükrözhet. A jogszabályok és felelősség kérdésében felmerül, hogy ki lesz a felelős egy önvezető autó döntéseiért, és hogyan kellene szabályozni az ilyen technológiákat, hogy minimalizáljuk az etikai problémákat.

Végül, a tudományos és társadalmi diskurzus során ezeket az eredményeket fel lehet használni az etika és a technológia kapcsolatáról szóló mélyebb megértés elősegítésére, hogy hogyan alakulhatnak ki azok a rendszerek, amelyek az etikai elveket követve működnek. Ez a kutatás rávilágít az emberek döntéshozatali folyamatainak komplexitására és ellentmondásosságára, különösen olyan helyzetekben, ahol a technológia befolyásolja a mindennapi életet.

4.1.3. Kulturális különbségek

A kulturális különbségek, miatt is más eredmény jött ki egyes országokban. Az egyén központubb országokban például (Egyesült Királyság, Franciaország, Egyesült Államok) nagyobb arányba mentenék meg a fiatalokat az idősekkel szemben, valamint több ember életét eggyel szemben még a nem egyén központú országokban (Kina, Japán) inkább az időseket mentené meg és nem tekintik központinak azt a kérdést, hogy hány ember élete forog kockán.

Japánba és az Egyesült Királyságban fontosabb a gyalogosok élete az utasokkal szemben, még Egyesült Államokban és Kínában ellenkezőleg gondolják ezt.

Szembe találkozhat olyan döntési esettel az önvezető autó, hogy szemben egy nagy tárgy lezuhanása miatt ki kell térnie jobbra vagy balra ahhoz, hogy elkerülje a balesetet. Amennyiben balra tér ki egy bukósisakot viselő motorost üt el, amennyiben jobbra úgy egy bukósisakot nem viselő motorost. Mi a helyes döntés ebben az esetben? Amennyiben balra tér ki az autó, hogy a lehető legkisebb következményt okozza, abban az esetben nem büntetés-e a motoros számára, hogy felelősen közlekedik? Viszont, ha jobbra tér ki az autó, abban az esetben meg megszegi az elvárást, hogy minimális következményt okozzon.

Felmerülnek a kérdések, hogy vannak, olyan életek, amik fontosabbak másokénál? A fenti kutatások alapján jöjjön létre, egy univerzális döntési módszer, ami nem minden kultúrába illeszkedik bele? Minden univerzális kérdés felírható lesz egy képlettel? Amennyiben nem születik egy univerzális megoldás, akkor a programozónak kell ezeket a kérdéseket, morális dilemmákat és életek kioltásáról szóló kérdéseket meghozni?

2018-ban Egyesült Államokbeli Arizonában egy önvezető Uber autó gázolt el egy nőt, aki végül bele is halt a sérüléseibe. Ez volt az első halálos eset, amit egy önvezető autó okozott és felmerült a felelősség kérdése mellett, technológiai kérdések is. Mivel előbbi már a dolgozat során kifejtésre került, ezért a következőkben az utóbbival foglalkozik a dolgozat.

4.1.4. Érzékelő rendszerek

Az önvezető autók érzékelőrendszerei különféle technológiákat kombinálnál a környezet pontos és biztonságos felismerése érdekében. Főbb érzékelő a LiDAR, Radar, Kamera, ultrahangos érzékelő és GPS.

A LiDAR egy infravörös vékony, láthatatlan lézerimpulzusokat kibocsátó optikai távérzékelő technológia, ami egy 3D-s térképet készít a környezetéről. Lézerimpulzusok sorát kibocsátva ad képet a tárgyak távolságáról és méretéről. A radar a rádióhullámok segítségével érzékeli az objektumok távolságát, sebességét és mozgási irányát. Kamerák azonosítják az útjelző táblákat, jelzőlámpákat, gyalogosokat és más járműveket. Ultrahangos érzékelő, egy rövid távolságú érzékelő, amelyet parkolás során használnak az autók. Ultrahangos hullámok visszaverődésével méri fel az autó körüli akadályokat. GPS pedig az autó földrajzi pontos meghatározását segíti.

Az önvezető autók különböző érzékelőrendszereit integrálni kell ahhoz, hogy a jármű képes legyen biztonságosan közlekedni és megfelelő döntéseket hozni. Az érzékelők által gyűjtött adatokat valós időben feldolgozzák és elemzik az autó fedélzeti számítógépei. Az érzékelők közötti redundancia és az adatok kombinálása növeli a rendszer megbízhatóságát és pontosságát.

Az önvezető autóknál 6 szint különböztethető meg a vezetés automatizálásától függően. A SAE J3016 szabvány definiálja az emberek és az autó rendszere közötti munkamegosztást.

A vezetőtámogató szintek az alábbiak:

- A 0. szinten egyáltalán nincs automatizmus, teljesen az ember kezében van az irányítás. Ezen autókban található automatikus vészfékezés, holttérfigyelés és sávelhagyás-figyelmeztetés.

- Az 1. szinten már bizonyos vezetéstámogatási funkciókkal rendelkező autók találhatók, de ezen funkciók mellett továbbra is az ember felelős az irányításért. A funkciók közé tartozik a sávközépen tartás **vagy** az adaptív tempómat.

- A 2. szinten részleges vezetésautomatizálási lehetőségek vannak az autóban. Az autó képes mindkét irányba manővereket végrehajtani, de emberi felügyeletet igényel. A funkciók közé tartozik a sávközépen tartás **és** az adaptív tempómat. A Tesla Autopilot rendszere is ebbe a kategóriába tartozik, mivel minden pillanatban jelezni kell, hogy a vezető részt vesz a vezetés folyamatában.

Az automatizált vezetési szolgáltatások:

- A 3. szinten az autó képes önállóan elvégezni a vezetés bizonyos részeit, de amennyiben az autó hibás működést tapasztal, a vezetőnek át kell vennie az irányítást a jármű felett. Ezen a szinten a legmagasabb automatizált funkció a forgalmi dugóban lévő automatizált vezetés.

- A 4. szinten az autó képes teljes mértékben önállóan közlekedni meghatározott körülmények között, például kijelölt városi területeken vagy autópályákon. Ezen a szinten már nem igényel sofőri jelenlétet, így a pedálok és a kormány sem feltétlen szükségesek az autóban. Ide tartoznak a vezető nélküli helyi taxik.

- Az 5. szinten teljes mértékben autonóm járművek találhatók, amelyek bármilyen közlekedési helyzetben és bármilyen útvonalon képesek önállóan közlekedni. Ezek az autók már minden körülmények között, a vezető beavatkozása nélkül képesek elvégezni a vezetési feladatokat.

Az egyes szintek közötti különbségek meghatározzák, hogy a jármű mennyire képes önállóan közlekedni, és milyen mértékben igényel emberi beavatkozást vagy felügyeletet. Az alacsonyabb szinteken (0-2) a vezetőnek folyamatosan figyelemmel kell kísérnie a vezetést, és szükség esetén be kell avatkoznia, míg a magasabb szinteken (3-5) az autó egyre inkább képes önállóan kezelni a közlekedési helyzeteket. (Barabás, 2017)

Az önvezető autók érzékelőrendszereinek fejlesztése és finomítása folyamatos kutatás és innováció tárgya, amelynek célja a közlekedés biztonságának és hatékonyságának növelése. Az érzékelők pontos és megbízható működése alapvető fontosságú ahhoz, hogy az önvezető autók elérjék a kívánt biztonsági szintet és társadalmi elfogadottságot.

Az önvezető autók rendszere is gépi tanulásra épül és a betáplált fotókból tanulja meg, hogy mi számít reklámtáblának, mi van az út szélén, milyen mikor egy gyalogos átmegy a zebrán stb., amennyiben a betáplált képek során egyes ember típusokból nem kap elég bemeneti adatot, nem fogja felismerni, hogy meg kell állnia az autónak, mert egy ember halad át a zebrán. Az adatok hiányossága mellett az éjszakai fényviszonyok miatt a sötétebb bőrű embereket nem minden esetben ismeri fel, akár infravörös világítás mellett sem a rendszer. A Gender Shades kutatás pont erre világítja rá a figyelmet.

4.1.5. Gender Shades kutatás

A kutatás (Buolamwini and Gebru, 2018) során a Microsoft, Face++ és az IBM mesterséges intelligencián alapuló arcfelismerő rendszerét tesztelték. Három Afrikai és három Európai országból 1270 képpel tesztelték a rendszereket, ebből 54,4 - 44,6 % a férfi - női arány és 46,4 - 53,6 % a sötétebb - világosabb bőrtónus aránya.

A nők- férfiak összehasonlításában mindegyik rendszer jobban teljesített a férfiak arcfelismerésében mint a nőkében, ez 8,1 - 20,6% -os hibaarányt jelent. A világosabb- sötétebb arcbőr közül a világosabbat ismerte fel a legtöbbször mind a három rendszer itt 11,8 - 19,2%-os hibaarány jött létre.

A legnagyobb különbség a sötétebb bőrű nők és a világosabb bőrű férfiak között alakult ki itt a hibaarány 34,4% volt. A Microsoft rendszer által téves rossz nemi csoportba soroltak közül 93,6%-ban sötétebb bőrű emberek voltak. Itt felvetül a kérdés, hogy a tanító adatok, nem voltak reprezentatívak egyes csoportokat tekintve, hiszen például a Microsoft esetében a világosabb bőrű férfiakat 100%-ban felismert a Face++ pedig a sötétebb bőrű férfiak esetében teljesített a legjobban 99,3%-ban. A tanulmány négy fontos okra hívja fel a figyelmet az önvezető autók tekintetében:

1. A tanulmány kimutatta, hogy a jelenlegi arcfelismerő rendszerek különböző hibaarányokkal működnek a különböző demográfiai csoportok esetében. Ez azt jelenti, hogy hasonló torzítások lehetnek jelen az önvezető autók által használt látás- és érzékelőrendszerekben is. Ha egy önvezető autó rendszeresen rosszul azonosítja vagy nem veszi észre a sötétebb bőrű embereket, az súlyos balesetekhez és igazságtalanságokhoz vezethet.

2. Az önvezető autók biztonságos működéséhez elengedhetetlen, hogy pontosan és megbízhatóan felismerjék a gyalogosokat, kerékpárosokat, járműveket és más akadályokat. A tanulmány eredményei rávilágítanak arra, hogy a jelenlegi technológiák még nem elég pontosak, különösen a különböző bőrtónusok és nemek felismerésében.

3. A tanulmány felhívja a figyelmet arra, hogy a tanító adatkészletek gyakran nem reprezentatívak. Az önvezető autók fejlesztéséhez használt adatoknak is tartalmazniuk kell különböző demográfiai csoportok képviselőit, hogy a rendszerek mindenkit egyformán jól felismerjenek és kezeljenek.

4. A tanulmány etikai kérdéseket is felvet például, hogy milyen előítéletek és diszkriminációk vannak jelen a technológiában. Az önvezető autók esetében ez különösen fontos, mivel ezek a járművek autonóm módon döntenek, amelyek emberi életeket érintenek. Az etikai szempontok figyelembevétele segíthet abban, hogy az önvezető autók döntései igazságosak és méltányosak legyenek.

Összegezve a tanulmány eredményeit rávilágít arra is, hogy hol vannak a jelenlegi technológiai hiányosságok, és milyen területeken van szükség további fejlesztésekre. Az arcfelismerő rendszerek pontosságának javítása és a torzítások csökkentése közvetlenül hozzájárulhat az önvezető autók érzékelőrendszereinek fejlesztéséhez.

4.2. Biztonsági protokollok

Az önvezető autó dilemmái közé sorolható, hogy feltörhetőek a rendszerek, így akár egy hacker feltörheti az autót, majd egy balesetet elkövetve a vezetőt vádolják meg, hogy ő a felelős a balesetért. Hogyan kerülhetőek el ezek a támadások?

A rendszereknek szükségesek többrétegű védelmet biztosítaniuk, amik magukba foglalják a tűzfalat, titkosítást és behatolásérzékelő rendszereket. Ezek segítségével figyelik az illetéktelen hálózati hozzáférést és szűrik a ki- és bemeneti forgalmat.

A redundancia biztosítja, hogy egy-egy komponens meghibásodása esetén a jármű továbbra is biztonságosan működjön, emiatt a kritikus alkatrészek például a szenzorok, vezérlő egységek duplikálva vannak a járműben ezzel elkerülve az esetleges rendszerleállásokat abban az esetben, ha az egyik alkatrész tönkremenne. Nem csak az alkatrészek, hanem a szoftverek is folyamatosan felülvizsgálják egymás döntéseit, ezzel elkerülve a hibás algoritmusok által hozott döntéseket.

A rendszerek folyamatos frissítése és hibajavítások gátolják, hogy az autók sebezhetőek legyen a támadásokkal szemben. Ezek mellett pedig gyakori a szimulált támadások és behatolási tesztek ezzel felmérve a rendszer gyengeségeit és sebezhetőségeit.

Ezek az intézkedések együttesen biztosítják, hogy az önvezető autók biztonságosak és megbízhatóak legyenek, védelmet nyújtva a rendszerhibák és kiberbiztonsági fenyegetések ellen.

5. COMPAS rendszer és adathalmaz

A COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) rendszer az Egyesült Államokban több tagállamban is használják a bűnügyi igazságszolgáltatásban. Az eszközt Northpointe, Inc. fejlesztette ki, de a forráskód nem publikus a nyilvánosság számára. Az eszköz célja, hogy az ügyészséget és a bírókat segítse a feltételes szabadlábra helyezés bizottsági döntésében és az egyes elkövetők, milyen integrációs folyamaton menjenek keresztül, hogy a lehető legjobban visszailleszkedjenek a társadalomba és ezáltal növeljék a közbiztonságot.

5.1. COMPAS adathalmaz és rendszer működése

COMPAS adathalmaz egy kérdőívre épül, ahol a letartoztatott személyek 137 kérdést válaszolnak meg. A kérdések között szerepelnek demográfiai, társadalmi, gazdasági háttér, korábbi bűncselekménnyel kapcsolatos, valamint pszichológiai állapotra irányuló is. A válaszokból, az alábbi képlet alapján létrejön egy rizikószint (alacsony, közepes, magas) aminek a jelentése, hogy milyen valószínűséggel, fog újabb bűncselekményt elkövetni a letartoztatott személy. Az erőszak visszaesésének kockázati értéke az alábbi képlettek jön ki:

(életkor $\cdot$ $(-w)$)+

(életkor első letartóztatáskor $\cdot$ $(-w)$)+

(szakképzés $\cdot$ w)+

(erőszakos múlt $\cdot$ w)+

(múltbeli fegyelmi eljárás $\cdot$ w)

A súly(w) az egy statisztikai módszerrel kiszámolt az adatgyűjtés során létrejött adatok közötti korrelációt jelenti. A rendszer által meghozott kockázati döntéseket, korábbi bírói döntéseken alapulnak, amik szintén lehetnek torzítottak, emiatt is a rendszer által meghozott eredményt a bírók, csak segítségnek használhatják és a végleges döntésben a bíróknak megfelelően kell indokolniuk, hogy milyen tényezőket vettek figyelembe a kockázatelemzés során. A COMPAS rendszer használata elött négy figyelmeztetést kell elfogadniuk a bíróknak:

1. Nem fedhetik fel, hogy milyen kockázati tényezőket vett figyelembe a rendszer.
2. A COMPAS rendszer nemzetközi adatokra épül, ezáltal nem reprezentatív egyes államokra koncentrálva.
3. Néhány tanulmány szerint a rendszer a kisebbségekkel szemben aranytalanul nagy kockázati értéket adhat.
4. A rendszert folyamatosan felülvizsgálni kell.

Az említett figyelmeztetések ellenére a COMPAS rendszer széles körben alkalmazott eszköz a büntető igazságszolgáltatásban, mivel számos előnnyel járhat a döntéshozatal folyamatában. Azonban az ilyen rendszerek használata felveti az etikai és jogi kérdéseket, különösen a diszkrimináció és az átláthatóság tekintetében.

5.2. A COMPAS rendszer előnyei és hátrányai

5.2.1. Előnyök

A rendszer előnyei közé sorolható, hogy a rendszer nagy mennyiségű adatfeldolgozására képes rövid időn belül és ez lehetővé teszik a bírók számára, hogy gyorsabb és hatékonyabb döntéseket hozhassanak meg.

A rendszerek kevésbé alkalmasak érzelmi és szubjektív befolyásoltság alá kerülni, emiatt objektív döntéseket hoz meg a rendszer. A rendszer statisztikai módszerekkel elemzi az adatokat, amely pontosabb és megalapozottabb döntéseket eredményezhet, különösen nagy adathalmazok esetén.

5.2.2. Hátrányok

A forráskód nem publikus, ami korlátozza a rendszer működésének és döntéshozatali logikájának átláthatóságát. Ez megnehezíti a rendszer megbízhatóságának és pontosságának ellenőrzését.

Tanulmányok kimutatták, hogy a rendszer kisebbségekkel szemben aránytalanul nagy kockázati értéket adhat. Ez a probléma abból adódhat, hogy a rendszer a múltbeli adatokon alapul, amelyek magukban hordozhatják a társadalmi és intézményi elfogultságokat.

Az automatizált döntéshozatali rendszerek esetében nehéz meghatározni, hogy ki vállalja a felelősséget a rendszer hibáiért vagy téves döntéseiért. Ez különösen fontos a büntető igazságszolgáltatásban, ahol a döntések jelentős hatással vannak az egyének életére.

5.3. A COMPAS rendszer kritikája: A ProPublica jelentése

A COMPAS rendszert számos kritika érte működését és eredményeit tekintve. Az egyik legátfogóbb és legmeghatározóbb elemzés a ProPublica nevű oknyomozó újságírói szervezethez köthető, amely 2016-ban jelentette meg tanulmányát a rendszerrel kapcsolatban. A ProPublica vizsgálata több fontos problémát azonosított a COMPAS rendszer használatával és működésével kapcsolatban.

A ProPublica elemzése rámutatott arra, hogy a COMPAS rendszer jelentős mértékben elfogult a sötétebb bőrű vádlottakkal szemben. Az adatok alapján a rendszer a sötétebb bőrű vádlottakat sokkal gyakrabban sorolta magas kockázati kategóriába, mint a világosabb bőrűek, annak ellenére, hogy később nem követtek el újabb bűncselekményt. Az elemzés kimutatta, hogy a sötétebb bőrű gyanúsított esetében kétszer nagyobb volt a valószínűsége annak, hogy a rendszer tévesen magas kockázatot jelöljön meg, mint a világosabb bőrű esetében. Ez az eredmény súlyos etikai és jogi kérdéseket vet fel a rendszer használatával kapcsolatban, különösen a faji egyenlőség szempontjából.

A rendszer pontosságát is megkérdőjelezte a tanulmány. A ProPublica szerint a rendszer által adott kockázati értékek csak mérsékelten voltak pontosak a valós visszaesési arányok előrejelzésében. Mind a sötétebb-, mind a világosabb bőrű vádlottak esetében hasonló arányban adtak téves előrejelzéseket, azonban a

sötétebb bőrű vádlottak esetében a téves pozitív előrejelzések aránya lényegesen magasabb volt. Ez azt jelenti, hogy a rendszer gyakran jelölte meg a sötétebb bőrűeket magas kockázatúként, amikor valójában nem jelentettek nagyobb visszaesési kockázatot. Ez azért lehetséges, mivel a rendszer tanuló adatbázisa korábbi ítéleteken alapszik, amik szintén torzítottak lehettek a sötétebb bőrű vádlottakkal szemben.

A COMPAS rendszer működésének átláthatósága is komoly kritikát kapott. A ProPublica jelentése szerint a rendszer zárt forráskódú, és a döntéshozatali folyamat részletei nem hozzáférhetők a nyilvánosság számára. Ez a hiányos átláthatóság megnehezíti a rendszer elfogulatlanságának és pontosságának független értékelését. Az ilyen zárt rendszerek esetében különösen fontos lenne, hogy a működésük és döntéshozatali logikájuk ellenőrizhető legyen, hogy elkerülhetőek legyenek az esetleges torzítások és hibák.

A ProPublica elemzése rámutatott arra is, hogy a COMPAS rendszer által használt kérdőív számos szociális és gazdasági tényezőt is figyelembe vesz, amelyek közvetetten hozzájárulhatnak a faji és társadalmi-gazdasági elfogultságokhoz. A kérdések között szerepelnek olyanok is, amelyek a vádlottak társadalmi-gazdasági helyzetére, például a munkanélküliségre vagy a lakhatási körülményekre vonatkoznak. Ezek a tényezők a rendszer döntéseit torzíthatják, és így hozzájárulhatnak a kisebbségi csoportok hátrányos megkülönböztetéséhez.

A ProPublica jelentése számos súlyos problémára hívta fel a figyelmet a COMPAS rendszerrel kapcsolatban. A faji elfogultság, a pontosság kérdései, az átláthatóság hiánya és a szociális-gazdasági tényezők torzító hatása mind olyan kihívások, amelyek sürgős megoldást igényelnek. Ahhoz, hogy a COMPAS rendszer valóban hozzájárulhasson a büntető igazságszolgáltatás hatékonyságának és igazságosságának növeléséhez, szükség van a rendszer működésének átfogó felülvizsgálatára és a szükséges reformok bevezetésére.

5.4. COMPAS adatelőkészítés

A hatékony mesterséges intelligencia rendszerek működéséhez elengedhetetlen az alapos adatfeldolgozás. A COMPAS adatbázis előkészítése során a következő módszereket és lépések hajtódtak végre.

5.4.1. COMPAS adat torzítottság

Az adatok előkészítése elött szükséges, volt felmérni az adatok minőségét és az esetleges torzítottsági lehetőségeket. A COMPAS rendszer által generált eredmények historikus adatokra épülnek, azaz régebbi bírók által hozott döntésekből tanul a rendszer. Az afroamerikaikat vagy más kisebbségi csoportokat aránytalanul gyakran állították meg, vizsgáltak át és tartoztatták le, mint egyéb más társadalmi csoportot. Ennek oka lehet a szisztematikus rasszizmus, profilalkotás vagy a rendőrségi erőforrások egyenlőtlen elosztása. Ha ezeket a torzított adatokat használják fel egy algoritmus betanítására, az algoritmus is ugyanezeket a torzításokat fogja tükrözni. Ilyen bizonyított eset a New Yorkban történt stop-and-frisk politika, aminek során a rendőrség aránytalanul gyakran állította meg és vizsgálta át az afroamerikai és latinó férfiakat. Ezek az események bekerülnek az adatbázisba, és azt sugallják, hogy ezek a csoportok nagyobb bűnözési hajlamot mutatnak, ami nem feltétlenül igaz.

A múltbeli döntések valóságtartalmát, és igazságosságát bizonyítani nem lehet, de kétségekkel kell kezelni. Az, hogy a döntés egy ember hozta meg nem jelenti azt, hogy nem tartalmaz torzítottságot, vagy diszkriminációt az ítélet. Ebből adódik az is, hogy a rendszer által javasolt értékek diszpropcionálisak lehetnek. Diszpropcionális döntéskor bizonyos csoportok rendszeresen súlyosabb büntetéseket kapnak ugyanazért a bűncselekményért, hiszen az algoritmus ezt a mintát tanulja meg, ha nincsen előfeldolgozva és kiegyenlítve a tanuló adatkészlet. Ez a mintázat tovább erősítheti a torzítottságot a modellben.

A kérdőív során felmerülnek társadalmi és gazdasági háttérre vonatkozó kérdések is. A munkanélküliség, iskolai végzettség vagy lakhatási helyzet nem feltétlen kapcsolódnak közvetlenül a visszaesési kockázathoz. A munkanélküliség az egyik visszaesési változóként szerepel az algoritmusban, azonban a munkanélküliség mértéke nem feltétlenül tükrözi közvetlenül az egyén hajlamát a bűncselekmények elkövetésére. Sok tényező befolyásolja a munkanélküliséget, például a gazdasági környezet, a diszkrimináció a munkaerőpiacon, valamint az oktatási és képzési lehetőségek hozzáférhetősége. Ha az algoritmus a munkanélküliséget figyelembe veszi, előfordulhat, hogy olyan egyének is magasabb kockázati besorolást kapnak, akik valójában nem hajlamosabbak a bűncselekmények elkövetésére, csak hátrányos helyzetben vannak a munkaerőpiacon.

A lakhatási helyzet során ugyanez az eset áll fent. Abban az esetben, ha valaki a szegényebb városrészben él, ahol magasabb a bűnözési ráta, nagyobb valószínűséggel kap magasabb kockázati besorolást, függetlenül attól, hogy ő maga

milyen valószínűséggel követne el bűncselekményt.

Az alacsony iskolai végzettség aránya szintén magasabb lehet olyan közösségekben, amelyekben korlátozottak az oktatási lehetőségek. Az ilyen változók figyelembevétele torzíthatja az eredményeket, mivel az oktatási egyenlőtlenségek nem feltétlenül tükrözik az egyén bűnözési hajlamát.

Az implicit előítéleteket akarva akaratlanul a fejlesztők is belecsempészhetik, amennyiben bizonyos változókra nagyobb súlyt helyeznek. A COMPAS esetében a kockázati szint kiszámításának során figyelembe veszik a szakképzést / iskolai végzettséget, ami ahogy fent is ki lett hangsúlyozva nem feltétlen befolyásolja a visszaesésé kockázatát.

A nem reprezentatív változókat szükséges kiszűrni, ezzel elősegítve, hogy az algoritmus torzítottság mentesen tudjon működni.

5.4.2. Adattisztítás és Előkészítés

A modellek célja, hogy torzitottság mentes adatokból képes legyen meghatározni a DecileScore-t, vagyis a visszaesés kockázatának a szintjét.

A ProPublica által közzétett adatbázis már előzetesen tisztított adatokat tartalmazott, amelyekben megtörtént a duplikációk szűrés, a kiugró adatok eltávolítása és az azonos oszlopok törlése.

Ezt követően a dátum oszlopok megfelelő formátummá alakítása történt meg, és a ”First name” és ”Last name” oszlopok összevonása, mivel a két adattábla ezen oszlopok mentén kapcsolódik össze. Az összekapcsolás során inner join módszert lett alkalmazva, ami azt jelenti, hogy csak a két tábla metszete került bele az új táblába.

Az összekapcsolt táblában megtörtént a pszeudonimizálás, vagyis állnevesítés a GDPR előírásainak betartása érdekében, és innentől kezdve a Person_ID alapján történik a megkülönböztetés. Az alábbi szűrési feltételek lettek alkalmazva:

- **days_b_screening_arrest** értéke -30 és 30 közötti legyen, mert egy hónapon túl a válaszok megváltozhatnak.
- **is_recid** értéke ne legyen -1, mert ez azt jelenti, hogy nem ismert, hogy az elkövető követett-e el újabb bűntényt.
- **c_charge_degree** értéke ne legyen ”O”, mert ezek általában jelentéktelen vádpontokat jelentenek.
- **score_text** értéke ne legyen ”N/A”, hogy csak olyan eseteket vegyünk figyelembe, amelyekhez tartozik kockázati pontszám.

Minden Person_ID-nak csak a legrégebbi screening_date -je lett megtartva elkerülve a duplikálást és az esetleges régebbi adatok megtartását. Ezek után az összes dátum oszlop törlésre került, hogy a modell folyamatot ne zavarja, hiszen ezeknek a változóknak nem szabad befolyásolnia az algoritmust.

Az eredi adatbázis tartalmazza a nem publikus modell eredményeit így először egy kimutatást végeztem, hogy az eredeti modell által meghatározott DecileScore csoportokban átlagosan milyen a visszaesésé aránya.

DecileScore	Átlag_is_recid
1	0.281176
2	0.416244
3	0.488591
4	0.540931
5	0.591497
6	0.611345
7	0.663912
8	0.777328
9	0.823529
10	0.778947

1. táblázat. Alap modell szerinti átlagos visszaesési arány DecileScore-okra lebontva

Az eredmény az mutatja, hogy az eloszlása a csoportoknak megfelel a kritériumoknak, ugyanis folyamatosan növekszik minél magasabb DecileScore kategóriába van beosztva kivéve a 10-es kategóriánál, de nincs olyan kategória, ahol 90%-os vagy magasabb arányban tudja megmondani, hogy visszaeső bűnöző-e. Emiatt is megfigyelésre került az etnikum alapú csoportokba való eloszlás:

DecileScore	1	2	3	4	5	6	7	8	9	10
African-American	278.0	369.0	390.0	352.0	318.0	309.0	263.0	184.0	163.0	72.0
Asian	10.0	4.0	1.0	4.0	0.0	2.0	0.0	1.0	0.0	0.0
Caucasian	361.0	275.0	265.0	194.0	147.0	115.0	72.0	42.0	39.0	15.0
Hispanic	114.0	80.0	55.0	39.0	52.0	30.0	14.0	14.0	11.0	5.0
Native American	2.0	1.0	0.0	3.0	0.0	0.0	1.0	0.0	2.0	0.0
Other	85.0	59.0	34.0	31.0	24.0	20.0	13.0	6.0	6.0	3.0

2. táblázat. Eredeti modell szerinti etnikum eloszlás a DecilScore alapján

Ez jól mutatja, hogy az afroamerikainak sokkal nagyobb arányban kerültek a magasabb kategóriákba, mint a többi etnikum, nyilván itt felmerül, azaz indok is, hogy sokkal magasabb számban található meg az afroamerikai etnikum az adathalmazban. Ezzel szemben a kaukázusi csoportoknál az alacsonyabb kategóriákban több elkövető található meg, míg a magasabb kategóriákba egyre kevesebb elkövetőt sorolt be a modell. Az ázsiai és hispán csoportok esetében szintén az alacsonyabb kategóriák dominálnak, és kevésbé jellemzőek a magasabb kategóriák. A bennszülött amerikai és egyéb csoportoknál az alacsony esetszám miatt nehéz konkrét következtetést levonni, de itt is inkább az alacsonyabb kategóriákban találhatóak az elkövetők. Ezek az eredmények rávilágítanak az egyes etnikumok közötti különbségekre a DecileScore eloszlásában.

5.4.3. Torzított adatok kiszűrése

A jellemzők a modell tanítás folyamatában játszanak fontos szerepet, hiszen ezek mentén fogja létrehozni a saját algoritmusát és döntési módszerét. A helyes jellemzők kiválasztása segít abban, hogy a modell elkerülje az esetlegesen nagy számú változókat, ami lassítaná a modell fejlesztését, tanítását és nagy mennyiségű rendszermemóriát használna fel. A jellemzők kiválasztása során eltávolításra kerül minden olyan bemeneti változó, ami nem informatív vagy esetleg redundáns adatokat tartalmaz. A felesleges változók okozhatják a modell romlását is.

Feltételezhető, hogy a tanító adat tartalmaz torzításokat, ezért szükséges, olyan jellemzők kiválasztása, ami minimálisra csökkenti a torzítottság lehetőségeit.

A torzítottság kiszűrése érdekében azon oszlopokat, amik torzított adatokat tartalmazhatnak kiszűrésre kerültek. A kiválasztott oszlopok a következőek lettek: kor, fiatalkori bűncselekmények száma, fiatalkori vétségek száma, egyéb fiatalkori bűncselekmények száma, korábbi bűncselekmények száma, napok a szűrés és a letartóztatás között, vád fokozata, vád leírása, napok a letartóztatástól számítva, bűncselekmény dátuma, erőszakos visszaesés, erőszakos bűncselekmény vádkának fokozata, erőszakos bűncselekmény dátuma, őrizetbe vétel dátuma, szabadlábra helyezés dátuma, esemény és etnikum.

5.4.4. Modell tanítás

A modell tanítása során az első lépés, hogy a string vagy object típusú adatokat átkonvertálni numerikus formátummá egy Label Encoder segítségével. Ez a lépés elengedhetetlen, mert a modellek, különösen a gépi tanulási algoritmusok, nem képesek közvetlenül kezelni a szöveges vagy objektum típusú adatokat. Az átkonvertált adatok biztosítják, hogy a modell numerikus formában kapja meg az információkat, amelyeket képes feldolgozni és elemezni.

Mivel feltételezhető, hogy a tanító adatok torzítottak ezért, felügyelet nélküli tanítási algoritmusok lesznek alkalmazva a következőkben.

A modell tanitás során első lépésben a kiválasztott jellemzők és a fő komponensek közötti kapcsolat lett megvizsgálva a PCA segitségével. A PCA hiperparamétereinek 2 dimenzió és 2 főkomponens lett beállitva.

0	1
-0.126033	age
0.170962	juv_fel_count
0.254508	juv_misd_count
0.200396	juv_other_count
0.365558	priors_count
0.079689	days_b_screening_arrest
-0.186849	c_charge_degree
0.139828	c_charge_desc
-0.519395	r_charge_degree
0.107848	r_days_from_arrest
0.021884	r_offense_date
-0.000000	violent_recid
-0.347269	vr_charge_degree
0.054677	vr_offense_date
0.006213	in_custody
-0.002526	out_custody
0.457093	event
-0.202639	race

3. táblázat. Főkomponenshez hozzájárulási aránya

Látszik az eredményeknél, hogy az **r_charge_degree** és a **race** is negativ irányba befolyásolja a főkomponenst azaz, ha az adott változó értéke növekszik, az a főkomponens irányába történő elmozdulás ellentétes lesz. A **event** és az **vr_offense_date** pedig pozitiv irányba vagyis, ha a változó értéke növekszik, az a főkomponens értékének növekedésével jár.

Első modell tanítás

Az első modell esetében egy fuzzy c-means (FCM) klaszterezési algoritmus lett feltanítva tanító adatbázison.

A hiperparaméter ebben az esetben k = 10, hiszen a cél, hogy az új modell a DecileScore-okat generálja újra más logika alapján és az eredményeket össze lehessen hasonlítani. A másik hiperparaméter a maximális iteráció száma, ami ebben az esetben 2, ugyanis magasabb iteráció során összevonta a klasztereket és nem lett volna 10 klaszter, ami összehasonlítás alapot ad a modellek között.

Első modell eredménye

Először megtörtént a modell sziluett értékének kiszámolása. A modell sziluett értéke : **-0.02**. Az érték azt jelenti, hogy a modell nem jól különíti el a klasztereket, és átfedések lehetnek a klaszterek között. Ez arra utal, hogy az adatok nem megfelelően csoportosíthatók a jelenlegi klaszterezési beállításokkal, és lehetséges, hogy a klaszterek közötti különbségek nem eléggé markánsak. Az alacsony vagy negatív sziluett érték arra figyelmeztet, hogy szükség lehet a klaszterek számának vagy a használt jellemzőknek az újragondolására.

A modell visszaesési és etnikum alapú kiértékelése ugyanazzal a módszerrel történik, mint az alapmodell esetében, hogy egyszerűbb legyen az összehasonlítás a későbbiekben.

Csoport	Átlag_is_recid	Javasolt_DecileScore
1	0.036255	1
2	1.000000	9/10
3	0.814433	7
4	0.385081	4
5	0.900000	8
6	1.000000	9/10
7	0.732970	5
8	0.275362	2
9	0.782609	6
10	0.315789	3

4. táblázat. Első modell szerinti átlagos visszaesési arány javasolt DecileScore-okra lebontva

A táblázatban a csoportok nem egyeznek meg az eredeti modell által értékelt csoportokkal, azaz az 1-es nem a legalacsonyabb veszélyeztetési kategória, és a 10-es nem a legnagyobb. A modell csak klaszterekre tudja bontani az adatokat, és nem képes pontosan meghatározni a kategóriákat, ami utólagos munkát igényel. A csoportokból jól látható, hogy két olyan csoport is van, ahol az átlagos visszaesési arány 1, ami nyilvánvalóan egy magasabb kategóriájú csoportba sorolható lenne, ha nem tíz csoport lenne a cél. A modell eredményeinél megfigyelhető, hogy a 4. és az 5. csoport között igen nagy az eltérés, ami arra utalhat, hogy a klaszterezésnél nem sikerült finom hangolni az egyes csoportok közötti határokat. Az ilyen eltérések felülvizsgálata és esetleges korrekciója szükséges a pontosabb eredmények elérése érdekében.

Javasolt_DecileScore	1	9/10	7	4	8	9/10	5	2	6	3
African-American	1123.0	55.0	0.0	71.0	8.0	1353.0	32.0	3.0	53.0	0.0
Asian	15.0	2.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0
Caucasian	812.0	43.0	4.0	29.0	13.0	600.0	16.0	1.0	6.0	1.0
Hispanic	245.0	7.0	2.0	4.0	3.0	146.0	5.0	0.0	1.0	1.0
Native American	5.0	1.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0
Other	168.0	3.0	6.0	2.0	2.0	100.0	0.0	0.0	0.0	0.0

5. táblázat. Első modell szerinti etnikum eloszlás a javasolt DecilScore alapján

A etnikum alapú csoport megoszlás során látható, hogy az afroamerikai csoportot is ugyanúgy sorolja alacsonyabb kategóriákba mint a többi csoportot, valamint a 9/10-es csoportba tényleg csak azon elkövetők kerültek, akik valóban visszaeső bűnözők lettek későbbiekben. A magas afroamerikai arány a 9/10-es csoportban azért szembe tűnő, mert alapból ebből a csoportból képviseltetik magukat a legtöbben, ez lehet rendőri túlkapás vagy általános bűnelkövetési adat.

Második modell tanítás

A második modell tanítása során K-közép algoritmus lett használva. A második modell a centroidok kezdeti helyzetét véletlenszerűen választja ki ebben az esetben, száz különböző inicializációt hajt végre és a legjobbakat választja ki, vagyis ahol a legkisebb az inertia. A klaszterszám tíz és maximum 1000 iteráción keresztül fut, hogy megtalálja a legjobb centroidot.

Második modell eredménye

A második modell sziluett értéke **0.111**. Ez az érték már pozitív, ami azt jelzi, hogy a klaszterek valamivel jobban elkülöníthetők, mint az első modell esetében. Azonban a 0,11 körüli érték még mindig azt mutatja, hogy a klaszterezés nem optimális, és lehetnek átfedések a klaszterek között. Ez arra utal, hogy bár a második modell jobb, mint az első, még mindig van hely a további javításra, például a jellemzők újraválasztásával vagy a klaszterek számának finomhangolásával.

Group	Mean_is_recid	Javasolt_DecileScore
1	0.814815	5
2	1.000000	6-10
3	1.000000	6-10
4	0.128450	2
5	0.096241	1
6	0.139059	3
7	1.000000	6-10
8	0.419162	4
9	1.000000	6-10
10	1.000000	6-10

6. táblázat. Második modell szerinti átlagos visszaesési arány javasolt DecileScore-okra lebontva

A modell eredménye alapján öt csoport kerülhetne összevonásra, valamint a maradék öt csoport eloszlása nem feltétlenül optimális a csoportok kialakításához. A javasolt_DecileScore-oknál megfigyelhető, hogy a 3-as és a 4-es csoport között elég nagy ugrás történik, így kimaradnak a közepes kockázatú csoportok. Mivel a modellek célja, hogy 1-10-ig DecileScore-okat határozzanak meg emiatt ez a módszer nem teljesít jobban az alap modellnél, így nem javasolt ezen paraméterekkel az alkalmazása.

Javasolt_DecileScore	5	6-10	6-10	2	1	3	6-10	4	6-10	6-10
African-American	89.0	489.0	33.0	591.0	263.0	378.0	473.0	70.0	175.0	137.0
Asian	0.0	3.0	0.0	2.0	6.0	8.0	1.0	0.0	0.0	2.0
Caucasian	14.0	215.0	18.0	257.0	250.0	375.0	202.0	72.0	65.0	57.0
Hispanic	2.0	51.0	3.0	61.0	71.0	130.0	49.0	19.0	12.0	16.0
Native American	0.0	0.0	1.0	0.0	3.0	4.0	0.0	0.0	0.0	1.0
Other	3.0	24.0	1.0	31.0	72.0	83.0	32.0	6.0	13.0	16.0

7. táblázat. Második modell szerinti átlagos visszaesési arány javasolt DecileScore-okra lebontva

A többi modellel együtt az etnikum alapú megoszlása megfelelőnek tekinthető, így következtethető, hogy a modell nem tesz különbséget az etnikumok között, viszont ahogy fent is említve volt nem javasolt ezen paraméterekkel a használata.

Innen a harmadik negyedik modellek újak és a következtetés is

Harmadik modell tanítás

Az adatok hatékony feldolgozása és elemzése céljából egy autoencoder neurális hálózat és a KMeans klaszterezési algoritmus kombináció lett alkalmazva. Az autoencoder segítségével a bemeneti adatok tömörítése történt meg, míg a K-közép algoritmus a tömörített adatok alapján végzett klaszterezést.

Az autoencoder többrétegű perceptron tipusú felépitéssel rendelkezik ebben az esetben és két rejtett réteget tartalmaz 128 és 64 neuronnal, amik ReLu aktivációs függvényt alkalmaznak. A legbelső bottleneck réteg 32 dimenziós, amely a tömöritett adatokat reprezentálja. A decoder réteg is két rejtett réteg segitségével állitja vissza az adatokat az eredeti dimenzióba, ahol sigmoid kimeneti réteg aktivációt használ. Az autoencoder MSE veszteségfügvénnyel és Adam optimalizációval tanult 50 epoch-on keresztül. A K-közép klaszterezés pedig a második modellhez hasonlóan történt.

Harmadik modell eredménye

A harmadik modell sziluett értéke **0.085**.

Group	Mean_is_recid	Javasolt_DecileScore
1	0.701149	4
2	1.000000	7-10
3	0.140534	3
4	0.093750	2
5	0.997886	6
6	1.000000	7-10
7	0.995536	5
8	1.000000	7-10
9	1.000000	7-10
10	0.068263	1

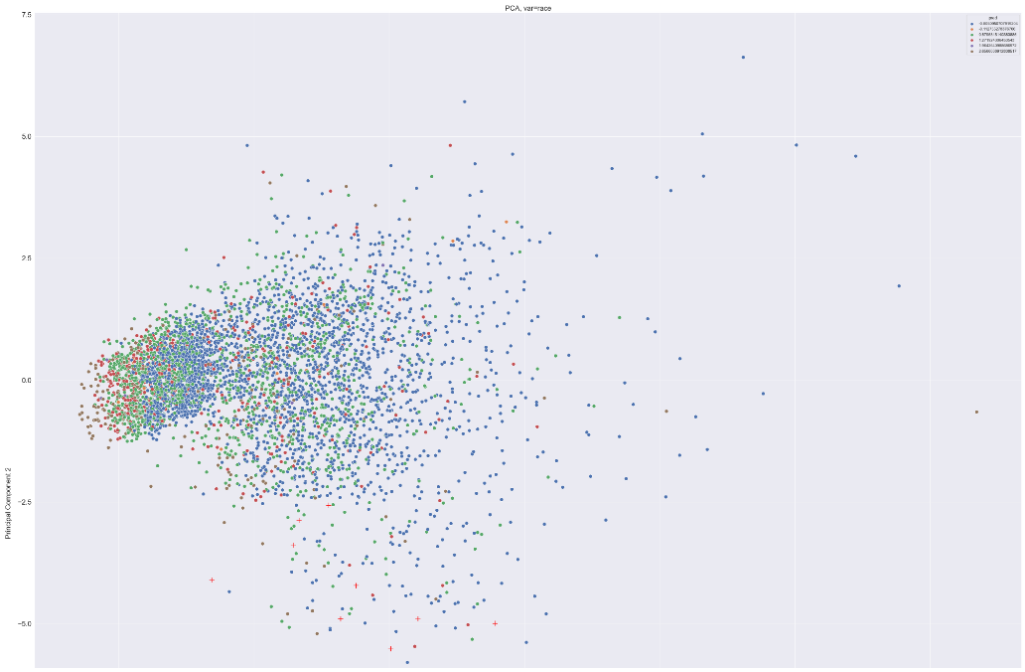
8. táblázat. Harmadik modell szerinti átlagos visszaesési arány javasolt DecileScore-okra lebontva

A klaszterek átlagos visszaesési aránya és a sziluett érték is jól mutatja, hogy nem megfelelő a klaszterezés, hiszen a klaszterek között nagy az átfedési arány igy négy olyan csoport is van ahol száz százalékos a visszaesési arány ami nem reális emiatt nem is használható a modell.

Javasolt_DecileScore	4	7-10	3	2	6	7-10	5	7-10	7-10	1
African-American	189.0	256.0	850.0	0.0	233.0	481.0	134.0	148.0	27.0	380.0
Asian	0.0	1.0	8.0	0.0	1.0	2.0	2.0	0.0	0.0	8.0
Caucasian	61.0	146.0	3.0	502.0	163.0	182.0	54.0	48.0	9.0	357.0
Hispanic	10.0	38.0	0.0	173.0	36.0	40.0	17.0	10.0	1.0	89.0
Native American	0.0	1.0	0.0	5.0	1.0	0.0	1.0	0.0	0.0	1.0
Other	1.0	17.0	0.0	184.0	39.0	9.0	16.0	12.0	3.0	0.0

9. táblázat. Harmadik modell szerinti átlagos visszaesési arány javasolt DecileScore-okra lebontva

A klaszterekben nem csak az átlagos visszaesési arány nem jól elkülönitett, hanem **race** alapján sem megfelelő, hiszen nem egyenletesek a csoportok közötti megoszlások egyik klaszterben sem.



8. ábra. PCA elemzés **race** alapján
(Saját szerkesztés)

A fenti grafikon egy két dimenziós főkomponens-analízis eredményeit mutatja. Az ábrán a pontok különböző színekkel jelölik a K-means által meghatározott klasztereket. A klaszterek célja, hogy az adathalmazt különböző csoportokba rendeződjenek, úgy hogy a legnagyobb legyen a csoportok között a variancia.

Az x-tengelyen az első fokomponens míg az y-tengelyen a második főkomponenst ábrázolja. Az ábrán a piros keresztekkel (’+’) vannak jelölve a klaszterek középpontjai, amiket a K-közép algoritmus határozott meg.

Az ábrán jól látható, hogy néhány klaszter szoroan összefügg, míg más klaszterek nagyobb mértékben elkülönülnek egymástól a térben.

Negyedik modell tanítás

A negyedik modell esetében az első és a harmadik modellt beállitásait alapul véve egy autoencoder neurális hálózat és az FCM klaszterezési algoritmus kombináció lett alkalmazva. A beállitások teljesen megegyeztek az autoencoder esetében a harmadik még az FCM algoritmus esetében az első modell beállításaival.

Negyedik modell eredménye

A negyedik modell sziluett értéke **0.042**.

Group	Mean_is_recid	Javasolt_DecileScore
1	0.070112	1
2	1.000000	7-10
3	0.997938	6
4	0.901042	3
5	0.993590	5
6	1.000000	7-10
7	1.000000	7-10
8	1.000000	7-10
9	0.707965	2
10	0.980831	4

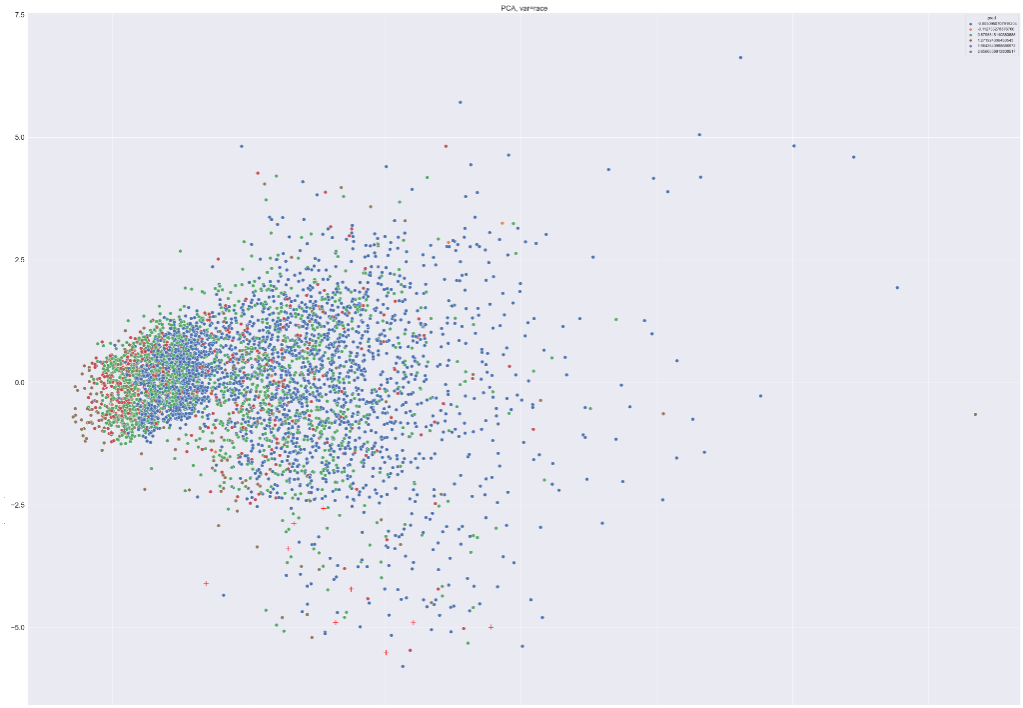
10. táblázat. Negyedik modell szerinti átlagos visszaesési arány javasolt DecileScore-okra lebontva

A harmadik modellhez hasonlóan a negyedik modell sziluett értéke és az átlagos visszaesési arány alapján a klaszterek nem megfelelően vannak kialakitva. Itt is négy száz százalékos átlagos visszaeséis arányú klaszter van ami nem megfelelő, valamint asziluett érték is jelzi, hogy nagy a klaszterek közötti lefedettség.

Javasolt_DecileScore	1	7-10	6	3	5	7-10	7-10	7-10	2	4
African-American	1199.0	82.0	234.0	127.0	129.0	641.0	9.0	2.0	99.0	176.0
Asian	15.0	2.0	3.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0
Caucasian	855.0	92.0	135.0	57.0	19.0	257.0	7.0	6.0	9.0	88.0
Hispanic	254.0	14.0	48.0	7.0	7.0	58.0	1.0	1.0	4.0	20.0
Native American	5.0	0.0	2.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
Other	168.0	1.0	63.0	1.0	1.0	15.0	0.0	3.0	1.0	28.0

11. táblázat. Negyedik modell szerinti átlagos visszaesési arány javasolt DecileScore-okra lebontva

A negyedik modell esetében a klaszterekben található **race** eloszlása jobb, mint a harmadik modell esetében, de mivel a fenti problémák továbbra is fent állnak emiatt nem alkalmas a modell.



9. ábra. PCA elemzés **race** alapján
(Saját szerkesztés)

A fenti grafikon a harmadik modell PCA elemzésével megegyezően két dimenziós főkomponens-analízis eredményeit mutatja. Az ábrán a pontok különböző színekkel jelölik az FCM által meghatározott klasztereket. A grafikonon jól látható, hogy nem elkülönültek és sokszor átfedések vannak a klaszterek között.

5.4.5. Következtetés

A négy vizsgált modell közül a K-közép modell bizonyult a leghatékonyabbnak az elítéltek elkülönítésében, különösen, ha kisebb módosításokkal biztosítjuk a visszaesési arányok megfelelő eloszlását. Az alapmodell kivételével kijelenthető, hogy a modellek nem tesznek különbséget az etnikai csoportok között, és nem használnak torzított adatokat, ami különösen fontos a méltányosság szempontjából.

Mivel az alapmodell kódja nem publikus, nincs lehetőség közvetlen kódszintű összehasonlításra. Az eredmények azonban egyértelműen azt mutatják, hogy a K-közép modell teljesít a legjobban az adatokon. Azonban ennek a modellnek is szüksége van további finomításokra, hogy éles környezetben is optimálisan működjön. Fontos lenne például a DecileScore értékének újragondolása annak érdekében, hogy valóban szükséges-e az elítélteket tíz csoportba osztani, vagy elegendő lenne három vagy öt csoportba sorolni őket. Ezzel a csoportosítás egyszerűsödhetne, és talán jobban tükrözné a valós kockázati szinteket.

További vizsgálatok és kísérletek szükségesek ahhoz, hogy a modell finomhangolásával biztosítsuk az optimális teljesítményt. Ehhez érdemes lenne bevonni a jogi és társadalomtudományi szakértőket is, hogy a modell ne csak technikailag, hanem társadalmilag is igazságos és elfogadható legyen. A modelleken szükséges módosítások végrehajtása után várhatóan jelentős előrelépést érhetünk el az etnikai alapú és visszaesési arányok eloszlásának javítása terén, ami a bűnmegelőzés hatékonyságának növeléséhez is hozzájárulhat.

Összességében, bár a modellek további finomításra szorulnak, az eddigi eredmények biztatóak, és lehetőséget kínálnak a méltányosabb és hatékonyabb ítéletelőrejelző rendszerek kifejlesztésére.

6. Összefoglalás

A kutatás elött és során is felmerült több kérdés és ezekre a primer kutatási módszerrel sikerült választ találnom. A kutatási kérdések:

1. Mit jelent, hogy torzított a mesterséges intelligencia?
2. Feltételezzük-e, hogy nem kell tökéletesnek lennie? Ha igen, elég csak annyi, hogy jobb mint az ember?
3. Hogyan lehet mérni a torzítottságot? Hogyan lehet elkerülni a torzítottságot?
4. Milyen következményekkel járhat a torzított mesterséges intelligencia
5. Miért jobb az, ha a mesterséges intelligencia végez el bizonyos döntéseket vagy számításokat az ember helyett?
6. Hogyan befolyásolja az önvezető autókat a torzítottság?

Az első kérdésnél egy definícióra kerestem a választ, amire, ahogy a mesterséges intelligenciára úgy a mesterséges intelligencia okozta torzítottságra sincsen egy univerzális definíció. A saját megfogalmazásomban az alábbi módon tudnám összefoglalni az MI okozta torzítottságot:

A torzított mesterséges intelligencia azt jelenti, hogy a mesterséges intelligenciai rendszerek olyan döntéseket vagy előrejelzéseket hoznak, amelyek szisztematikusan kedveznek vagy hátrányos helyzetbe hoznak bizonyos csoportokat vagy egyéneket. Ez a torzítás több különböző forrásból eredhet, és számos formában jelenhet meg. Amennyiben a kimeneti adatok nem kerülnek vizsgálatra, úgy ezen torzítottságok napvilágra sem feltétlen kerülnek.

A második kérdés során az MI tökéletessége és az emberrel szembeni teljesítménye merült fel. A véleményem szerint a mesterséges intelligenciának nem kell tökéletesnek lennie ahhoz, hogy hasznos legyen. Az emberi döntéshozatal és teljesítmény sem tökéletes, és a mesterséges intelligencia célja gyakran az, hogy ezen javítson. A tökéletesség elérése jelenleg technológiailag és etikai szempontból is kihívást jelent. A mesterséges intelligencia fejlődésével azonban a cél inkább az, hogy folyamatosan javuljon, minél több hibát kiküszöböljön, az adatfeldolgozás gyorsítása és olyan összefüggéseket vegyen észre, amire jelenleg az ember nem képes.

A harmadik kérdéskörben a torzítottság mérése és kiküszöböléséről volt szó. A torzítottság kiszűrésére, azonban nincs egyelőre egy tökéletes metódus. Statisztikai és matematikai szisztémákkal azonban van lehetőség ezek minimális kiszűrése, de ezen kutatási terület kialakulására és fejlődésére szükség van a jövőben. A torzítottság elkerülése már egy egyszerűbb kérdés, hiszen nagyrészben megvannak azok a kritikus pontok, ahol kialakulhatnak és befolyásolhatják a torzítottság mértékét. Amennyiben ezen eljárások betartása megtörténik, úgy elkerülhető a nagymértékű torzítottság.

A torzított mesterséges intelligencia következményei a negyedik kérdésben merültek fel. Amennyiben, a modell torzított végeredményeket ad, úgy akár egyes csoportok előnyben vagy hátrányba kerülhetnek a többi csoporttal szemben. Ezek a hatások nem csak az egyénre, hanem akár a társadalomra és a gazdaságra is hatással lehet. Az egyén szempontjából, egy hitel kérelem, szabadlábra helyezés vagy akár az önvezető autók által hozott döntés az életük további részére is hatással lehet. Amennyiben ezek a döntések a társadalom egyes csoportjaira hat, abban az esetben ezen csoportok irányába más területen is nőni fog a diszkrimináció.

Az ötödik kérdés a dolgozat írása közben merült fel, hogy miért jobb, ha az MI végez el döntéseket, számításokat az ember helyett. A mesterséges intelligencia előnyeihez sorolható, hogy hatalmas mennyiségű adatot képes feldolgozni igen rövid idő alatt, kisebb hibaszázalékkal a nap huszonnégy órájában. Ezen faktorok mellett az MI-ban nincsenek előítéletek és nem befolyásolják az érzelmei egyes döntések meghozatalakor. Ezek mellett a mesterséges intelligenciát nem lehet megvesztegetni, megzsarolni és megfenyegetni azaz nem befolyásolható úgy, mint az ember.

A hatodik kérdés során fókuszban az önvezető autóknál megfigyelhető torzí-

tottság áll. Az önvezető autókat is ugyanúgy befolyásolja, hogy a tanító adatok nem adekvát vannak kiválasztva és ez hatást gyakorol akár az egyes bőrtípusok felismeréséné. Ez a torzítottság, akár emberi életekbe is kerülhet, ami miatt meglátásom szerint nem engedélyezném az önvezető autókat az utakon, addig amíg a morális és jogi kérdések tisztázásra nem kerülnek, valamint a rendszerek nem lesznek biztonságosak mindenki számára.

Összefoglalva a felvetődő kérdésekre nagyrészben választ kaptam, de vannak olyan dolgok, amiknek tisztázása szükséges ahhoz, hogy átlátható, biztonságos és igazságos alkalmazások és rendszerek alakuljanak ki. A jogi és etikai kérdések mellett a felelősök körét is szükséges lenne meghatározni, valamint a kritériumokat, hogy miknek kell megfelelnie a programkódnak, hogy minimális legyen a torzítottság egyes csoportokkal szemben. A cél, hogy a mesterséges intelligencia valóban segítse az emberiséget konzisztens és diszkriminativ mentes döntések meghozatalában.

7. Irodalomjegyzék

Alelyani (2021). Detection and evaluation of machine learning bias. Letöltve: 2024.05.23.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., and Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729):59–64. Retrieved from Nature.

Barabás, I. (2017). Current challenges in autonomous driving.

BME, M. 1.3 the foundations of artificial intelligence. http://project.mit.bme.hu/mi_almanach/books/aima/ch01s03. Accessed: 2024-08-07.

BME, M. 6.8 heuristic functions. http://project.mit.bme.hu/mi_almanach/books/aima/ch06s08. Accessed: 2024-08-07.

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 35.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 77–91. MIT Media Lab.

European Parlament (2024). Eu ai act: first regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

Európai Bizottság (2020). Fehér könyv a mesterséges intelligenciáról. https://commission.europa.eu/system/files/2020-03/commission-white-paper-artificial-intelligence-feb2020_hu.pdf.

Európai Parlament (2023). Az első uniós rendelet a mesterséges intelligenciáról. https://www.europarl.europa.eu/topics/hu/article/20230601STO93804/az-elso-unios-rendelet-a-mesterseges-intelligenciarol.

Forum, W. E. (2020). The future of jobs report 2020. "https://www.weforum.org/publications/the-future-of-jobs-report-2020/".

Fry, H. (2018). *Hello World : How to be Human in the Age of the Machine*. W.W. Norton & Company, New York.

Gordon, Glazner, Rawson, and Sturtz (2018). Reflections on the trolley problem. *Nature*, 563 :669–673.

Greene, D., J., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537) :2105–2108.

Hadke, B., Ingle, J., and Shetty (2021). Efficient clustering for weather forecasting using big data.

LeDoux, J. (1998). *The Emotional Brain : The Mysterious Underpinnings of Emotional Life*. Simon & Schuster.

Lexiq.hu (2024). Mesterséges neurális hálózat.

Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2022). A survey on bias and fairness in machine learning. Letöltve : 2024.05.23.

Moral Machine (2024). Moral machine. Accessed : 2024-08-09.

Pancake, T. (2022). Reflections on the trolley problem.

Pődör, L. (2020). Az önvezető járművek, a trolley probléma és az emberi élet védelme – széljegyzetek egy jogi-erkölcsi dilemma margójára. Letöltve : 2024.07.12.

Rudin, Wanf, C. (2020). The age of secrecy and unfairness in recidivism prediction. Letöltve : 2024.05.23.

Sharma (2024). K-means clustering explained. https://neptune.ai/blog/k-means-clustering.

Simon, H. A. (1980). *The Sciences of the Artificial*. MIT Press, 2nd edition.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236) :433–460.

Török, Z. (2021). A mesterséges intelligencia szabályozási kihívásai.

Weckert, S. (2020). Google maps hacks. https://www.simonweckert.com/googlemapshacks.html.

Wiki, P. (n.d.). Deep autoencoder. Accessed : 15 September 2024.

Ábrák jegyzéke

Táblázatok jegyzéke