Sandro Radovanović*

Univerzitet u Beogradu - Fakultet organizacionih nauka, Srbija

Andrija Petrović

Univerzitet u Beogradu - Fakultet organizacionih nauka, Srbija

Boris Delibašić

Univerzitet u Beogradu - Fakultet organizacionih nauka, Srbija

Milija Suknović

Univerzitet u Beogradu - Fakultet organizacionih nauka, Srbija

*sandro.radovanovic@fon.bg.ac.rs

PRAVEDNOST I JEDNAKOST U ALGORITAMSKOM ODLUČIVANJU

Apstrakt *U poglavlju se istražuje problem algoritamske pravednosti, posebno u kontekstu automatizovanih sistema odlučivanja i algoritama mašinskog učenja koji se sve više koriste u oblastima poput zapošljavanja, pravosuđa i obrazovanja. Kroz analizu najpoznatijih primera nepravednih algoritama pokazano je kako nepravedni podaci i modeli mogu dovesti do diskriminatornih odluka. U poglavlju je, takođe, prikazan obuhvatan pregled literature za rešavanje algoritamske nepravednosti, uključujući početnu pripremu podataka, prilagođavanje algoritama i naknadnu obradu predikcija. Rad prikazuje najbitnije rezultate koje je grupa autora ostvarila u ovoj oblasti. Eksperimentalni rezultati pokazuju da postoji neizbežan kompromis između pravednosti i tačnosti nevezano od pristupa rešavanja, ali se pokazalo da blago smanjenje tačnosti može značajno unaprediti pravednost, posebno na nivou grupe. Ukazuje se da, ipak, postoji potreba za daljim istraživanjem metoda koje kombinuju matematičke i društvene aspekte pravednosti, kao i za razvojem algoritamskih sistema koji će omogućiti transparentnost i odgovornost u donošenju odluka.*

Ključne reči**:** *Algoritamska pravednost, algoritamsko odlučivanje, mašinsko učenje, priprema podataka, prilagođavanje algoritama, nakadna obrada predviđanja.*

Uvod

U eri sve većeg oslanjanja na automatizovane sisteme i mašinsko učenje, algoritamska pravednost postaje ključni koncept u diskusijama o odgovornom razvoju tehnologije [4]. Algoritmi sve češće igraju presudnu ulogu u donošenju odluka u raznim oblastima, uključujući zapošljavanje [11], zdravstvenu zaštitu, pravosudni sistem [57], obrazovanje [25] i druge sektore od javnog značaja. Međutim, dok algoritamski kreirani modeli odlučivanja omogućavaju brže, efikasnije i objektivnije donošenje odluka, često nose sa sobom rizik reprodukovanja postojećih društvenih nepravdi i diskriminacije.

Ovaj rad ima za cilj da pruži uvod u problem algoritamske pravednosti, da objasni zašto je taj problem interesantan i bitan, kao i da istakne zbog čega je njegovo rešavanje kompleksno i zašto naivni pristupi, poput isključivanja osobina za koje se smatra da su osetljivi i po kojima ne bi trebalo da se donosi odluka, često nisu dovoljni za postizanje pravičnih rezultata. Potom, biće prikazani pristupi i načini rešavanja problema algoritamske pravednosti u oblasti višekriterijumske analize odlučivanja i metoda za višekriterijumsku analizu odlučivanja, kao i u algoritmima mašinskog učenja.

Algoritamska pravednost prvenstveno se odnosi na pitanje kako osigurati da automatizovani sistemi donose pravedne odluke, bez favorizovanja ili diskriminacije određene grupe ljudi. U širem smislu, problem algoritamske pravednosti postavlja pitanje kako možemo dizajnirati algoritme koji ne stvaraju, niti pogoršavaju postojeće oblike nejednakosti u društvu.

Na primer, u sistemima zapošljavanja, algoritmi koji procenjuju kandidate za posao mogu nesvesno favorizovati kandidate iz određenih etničkih ili socijalnih grupa, ako su obučeni na podacima koji odražavaju istorijsku diskriminaciju [11]. Slično, algoritmi koji se koriste u pravosudnom sistemu mogu davati pristrasne procene rizika za određene manjine, čime se potencijalno pogoršava njihov položaj u društvu [54].

Pravičnost u algoritamskim sistemima može se definisati na različite načine [2, 6, 3, 9], ali neka od ključnih pitanja uključuju:

- Pravednost u raspodeli resursa [52, 4, 31]: Da li algoritam jednako tretira sve grupe u društvu?
- Proceduralnu pravednost [10]: Da li su procesi koji vode do donošenja odluka transparentni i pošteni?
- Pravna i moralna odgovornost [3, 4]: Ko je odgovoran za posledice koje proizlaze iz upotrebe algoritama?

Pitanje algoritamske pravednosti je važno jer se algoritmi sve više koriste za donošenje odluka koje direktno utiču na živote ljudi. Ako su algoritmi nepravedni ili pristrasni, oni mogu imati ozbiljne posledice po pojedince i društvo u celini. Na primer, diskriminatorni algoritam u zapošljavanju može uskraćivati mogućnosti za određene grupe ljudi, dok pristrasan algoritam u pravosuđu može dovesti do nepravednih presuda. [31, 4]

Ono što ovaj problem čini posebno interesantnim jeste činjenica da su algoritmi često percipirani kao objektivni i nepristrasni alati, upravo zbog njihove osnove u matematičkim i statističkim modelima [6]. Međutim, ti modeli su obučeni na podacima koji su prikupljeni u svetu koji nije lišen nejednakosti i predrasuda. Time, algoritmi mogu nesvesno učvrstiti postojeće obrasce diskriminacije, umesto da ih eliminišu [4, 1].

Pored toga, algoritmi su izuzetno skalabilni, što znači da nepravedne odluke mogu uticati na veliki broj ljudi u veoma kratkom vremenskom periodu. Odnosno, kada bi pojedinac ili grupa samostalno donosili odluke ne bi bilo u mogućnosti da donesu toliko odluka i utiču na živote ljudi koliko može automatizovani sistem. Dakle, problem algoritamske pravednosti nije samo tehničko pitanje, već je i duboko društveno, pravno i etičko pitanje. [6, 1]

Iako je lako uočiti zašto je algoritamska pravednost bitna, rešavanje ovog problema predstavlja izuzetno veliki izazov iz nekoliko razloga.

Pravičnost je složen koncept [3], a njena definicija zavisi od konteksta i vrednosti društva. Tačnije, pravednost je individualni koncept koji se oblikuje kroz vaspitanje, školstvo, društvo i lično iskustvo [1]. Međutim, postoji više različitih formalnih definicija pravičnosti u okviru algoritama, poput jednakosti šansi, grupne pravičnosti i individualne pravičnosti. Međutim, često se dešava da se ove definicije međusobno isključuju. Na primer, algoritam može biti dizajniran da osigura jednakost šansi između različitih grupa, ali to može rezultirati nejednakim ishodima za pojedince [33]. Zbog toga, donošenje odluka o tome koja vrsta pravičnosti je najvažnija za konkretan kontekst može biti veoma teško.

Algoritamski modeli odlučivanja uče iz podataka koji su prikupljeni u realnom svetu. Ako su ti podaci pristrasni, algoritam će verovatno reflektovati tu pristrasnost u svojim odlukama. Na primer, ako istorijski podaci pokazuju da su manjinske grupe bile manje zaposlene, algoritam obučavan na takvim podacima može nastaviti da favorizuje većinsku grupu [47]. Dakle, čak i najnapredniji algoritmi nisu imuni na reprodukciju nepravdi, jer su podaci na kojima se obučavaju oblikovani društvenim obrascima i pristrasnostima [4].

Mnogi algoritmi, posebno oni zasnovani na metodama dubokog učenja, funkcionišu kao crne kutije, što znači da je teško razumeti kako dolaze do svojih odluka [1, 9]. Ovo otežava identifikaciju i ispravljanje nepravičnih odluka. Međutim, to ne znači da se problem pravednosti jednostavno ne može objasniti i da se njime ne treba baviti. Naprotiv, odsustvo razumevanja kako je odluka doneta ne lišava odgovornosti lica koje koristi algoritamske sisteme od odgovornosti odluke [53]. Povezano sa algoritmima crnih kutija jeste i nedostatak transparentnosti koja može dodatno komplikovati situacije u kojima je potrebno objasniti ili opravdati određene odluke koje algoritam donosi [53]. Nažalost, zakonodavstvo i standardi za algoritamsko odlučivanje još uvek nisu dovoljno razvijeni. Postoji manjak pravnih okvira koji regulišu upotrebu algoritama u odlučivanju (ili su puno nejasnoća i nekompletnosti), što može dovesti do situacija u kojima organizacije implementiraju nepravedne algoritme bez posledica [18, 56, 19].

Sa druge strane, rešavanje problema algoritamske pravednosti je teško zato što naivni pristupi nisu dovoljni. Na primer, jednostavno uklanjanje varijabli poput rase ili pola iz podataka na kojima se algoritmi obučavaju ne garantuje pravičnost [4]. Razlog za to je što mnogi drugi atributi mogu biti povezani sa ovim varijablama i mogu indirektno doprineti diskriminaciji. Ova pojava je poznata kao posredna pristrasnost, gde se pristrasnost i dalje manifestuje kroz druge podatke koji su povezani sa demografskim karakteristikama.

Drugi naivni pristup bio bi korišćenje algoritama koji optimizuju samo za tačnost, bez uzimanja u obzir pitanja pravičnosti. Međutim, algoritmi mogu biti veoma tačni, ali i dalje nepravedni prema određenim grupama. Tačnost sama po sebi nije dovoljna metrika za evaluaciju algoritama koji imaju društveni uticaj. Još jedan naivan pristup bio bi pokušaj balansiranja pravičnosti jednostavnim metodama poput forsiranja jednakih stopa grešaka između različitih grupa. Iako ovo može izgledati kao intuitivno rešenje, često dolazi do značajnih kompromisa u tačnosti modela, što može dovesti do drugih vrsta nepravdi [33, 1].

U narednim poglavljima biće opisani istaknuti primeri nepravednosti algoritamskog odlučivanja kako bi čitaoci bili upoznati sa posledicama u realnim sistemima. Nakon toga biće opisan pregled literature, odnosno kako se meri nepravednost i koji su pristupi dostupni u literaturi. Potom, opisaće se pristupi u rešavanju nepravednosti grupe autora koji se mogu kategorizovati u metode pretprocesiranja podataka, promene algoritma i postprocesiranja rezultata u cilju ostvarivanja pravednosti, te i razvoj pravednih modela višekriterijumske analize odlučivanja i analize obavijanja podataka. Takođe,

posvetiće se pažnja oko neusklađenosti mera pravednosti sa željenim rezultatom.

Istaknuti primeri nepravednosti algoritamskog odlučivanja

Kako bismo razumeli sam problem (ne)pravednosti i (ne)jednakost u algoritamskom odlučivanju potrebno je da opišemo par primera. Ispod će biti opisani najistaknutiji primeri nepravednosti u algoritamskom odlučivanju.

Amazonov algoritam za zapošljavanje

Amazonov problem s algoritmom za zapošljavanje predstavlja jedan od najpoznatijih primera kako automatizovani sistemi mogu nenamerno algoritamski pojačati pristrasnost koja je prisutna u podacima. Početkom 2014. godine, Amazon je počeo da razvija alat za rangiranje kandidata sa ciljem ubrzavanja procesa zapošljavanja koristeći algoritme mašinskog učenja da oceni potencijalne zaposlene na osnovu njihovih biografija [58]. Ideja je bila da se algoritamski sistem osposobi da iz velikog broja prijava izdvoji najbolje kandidate, oslanjajući se na istorijske podatke o zaposlenima i kriterijume koje su kompanijski regruteri koristili prilikom donošenja odluka [11].

Algoritamski model odlučivanja je bio obučavan na podacima prikupljenim iz biografija kandidata koji su se prijavljivali na tehničke pozicije u prethodnih 10 godina. Većina biografija pripadala je muškim kandidatima, što je odražavalo širi društveni problem u industriji tehnologije, gde su muškarci često veoma zastupljeniji u tehničkim ulogama. Kao rezultat toga, algoritam je „naučio" da preferira biografije koje su bile slične muškim kandidatima, penalizujući kandidate čije su biografije sadržavale reči povezane sa ženama. Na primer, biografije koje su sadržavale fraze poput „ženski tim za fudbal" bile su automatski ocenjene lošije [11]. Posledično, procenat žena koji su se javljali kao najbolji kandidati se smanji dovodeći do sistematske diskriminacije žena, iako sama polna varijabla nikada nije eksplicitno uneta u model [11, 34].

Nakon što su uočili problem, inženjeri u Amazonu pokušali su da uklone eksplicitne reference na pol iz algoritma, u pokušaju da spreče ovu pristrasnost. Međutim, to nije rešilo problem. Algoritam je i dalje nalazio indirektne načine da penalizuje ženske kandidate, oslanjajući se na druge varijable koje su asocirale na pol kandidata, kao što su obrazovni putevi [34]. Ovaj fenomen poznat je kao posredna pristrasnost, gde se algoritmi oslanjaju na varijable koje služe kao zastupnici za diskriminisane osobine, čak i kada su te osobine direktno uklonjene. U ovom slučaju, iako pol nije bio direktno

uključen u podatke, druge karakteristike su bile dovoljno povezane s polom da omoguće algoritmu da nastavi s diskriminacijom. Amazon je pokušao da na razne načine unapredi algoritam, ali problem je bio dublji. Naivni pristupi, poput jednostavnog uklanjanja varijabli povezanih s polom ili forsiranja jednakog tretmana između različitih grupa, nisu dali rezultate [11, 48]. Modeli algoritamskog odlučivanja su, zbog prirode podataka i njihovih obrazaca, nastavili da favorizuju muške kandidate. Ovaj problem ukazuje na ključni izazov u oblasti algoritamske pravednosti: jednostavno uklanjanje diskriminatornih varijabli ili forsiranje matematički jednakih rezultata ne garantuje pravičnost. Algoritmi su u svojoj suštini odraz podataka na kojima su obučavani, a ako ti podaci nose sa sobom istorijske pristrasnosti, algoritmi će nesvesno reflektovati te pristrasnosti [48, 47].

COMPAS

COMPAS (eng. *Correctional Offender Management Profiling for Alternative Sanctions*) je softverski alat koji se koristio u američkom pravosudnom sistemu za procenu rizika od recidiva, tj. verovatnoće da će optuženi ponovo počiniti krivično delo [57, 28]. Ovaj alat je razvijen kako bi pomogao sudijama u donošenju odluka o kauciji, uslovnoj kazni i drugim aspektima krivičnog pravosuđa. Iako je COMPAS dizajniran da bude objektivno sredstvo za procenu rizika, postao je predmet ozbiljne kritike zbog nepravednosti prema etničkim manjinama, posebno prema ljudima tamnije boje kože [28].

Problem sa COMPAS sistemom prvi je postao široko poznat nakon što je organizacija ProPublica 2016. godine sprovela istraživanje koje je ukazalo na ozbiljne pristrasnosti u ovom algoritmu. Istraživanje je pokazalo da je algoritam sistematski davao veće ocene rizika ljudima tamnije boje kože, čak i kada su imali slične kriminalne dosijee kao ljudi bele boje kože. Analiza je pokazala da je COMPAS dvostruko češće pogrešno klasifikovao ljude tamnije boje kože kao visokorizične (tj. kao one koji će verovatno ponovo počiniti krivično delo), dok je belcima češće davao niže ocene rizika, čak i kada je postojala velika verovatnoća da će ponoviti krivično delo. Drugim rečima, COMPAS je bio pristrasan prema ljudima tamnije boje kože, sugerišući da će oni verovatno počiniti nova krivična dela u mnogo većem broju nego što je to zapravo bio slučaj. S druge strane, kod ljudi bele boje kože algoritamski model odlučivanja je bio sklon da potceni rizik od recidiva, što znači da su pojedini ljudi bele boje kože koji su imali visok rizik od ponovnog prekršaja ocenjivani kao niskorizični. [57, 16, 8]

Pristrasnost u COMPAS alatu nije proizašla iz toga što je algoritam eksplicitno uzimao u obzir rasu prilikom donošenja odluka, već iz pristrasnih podataka na kojima je obučavan [16, 28]. Kao i mnogi drugi algoritmi, COMPAS se oslanjao na istorijske podatke iz američkog krivičnog pravosuđa, koje već decenijama pokazuje duboke nejednakosti i pristrasnost prema manjinama. Ljudi tamnije boje kože su nesrazmerno zastupljeni u američkom zatvorskom sistemu i često su suočeni sa strožim kaznama i višim stopama hapšenja, čak i za ista krivična dela koja počine pripadnici drugih grupa. Ovi podaci oblikovali su COMPAS i njegovu percepciju rizika, što je dovelo do toga da algoritam reflektuje i pojačava istorijsku nepravdu [57, 28].

Još jedan ključni razlog za pristrasnost leži u načinu na koji je model postavljen. COMPAS koristi niz varijabli kako bi predvideo verovatnoću recidiva, uključujući demografske podatke, istoriju hapšenja, socioekonomski status i druge faktore. Iako rasa nije eksplicitno uzeta u obzir, mnogi od tih faktora su indirektno povezani s rasom zbog socijalnih i sistemskih nejednakosti u društvu. Na primer, stopa nezaposlenosti ili nivo obrazovanja mogu biti različiti među različitim rasnim grupama, i ti faktori mogu nesvesno dovesti do diskriminatornih rezultata [57, 28].

COMPAS je bio predmet brojnih pravnih i etičkih debata. Kritičari su istakli da algoritmi koji se koriste za donošenje odluka u pravosuđu moraju biti pravični, transparentni i odgovorni, jer direktno utiču na živote pojedinaca. Odnosno, jedan od glavnih problema sa COMPAS-om bio je i nedostatak transparentnosti. Model odlučivanja nije bio dostupan za javnu proveru, jer je bio vlasnički softver koji je razvila privatna kompanija, što je onemogućilo sudove, advokate i istraživače da u potpunosti analiziraju kako algoritamski model donosi odluke i zašto je odluka pristrasna [57].

Upis na fakultet u Velikoj Britaniji

Usled pandemije COVID-19 britanski studenti su bili sprečeni da polažu završne ispite (A-levels) koji su ključni za prijem na univerzitete [5]. U cilju da pokuša da obrazovanje održi na uglednom nivou britanska vlada i obrazovne institucije odlučile su da koriste algoritamske modele kako bi procenile njihove konačne ocene [25, 12]. Cilj je bio da algoritam generiše ocene na osnovu niza faktora, uključujući prethodne školske rezultate, ocene nastavnika i istorijske podatke o performansama škola. Nažalost, ovaj pokušaj je rezultovao u „*A-level* algoritamskoj aferi“, koji je izazvao široku javnu polemiku o algoritamskoj pravednosti i o tome kako se automatizovani sistemi

koriste za donošenje važnih odluka koje direktno utiču na živote mladih ljudi [30].

Umesto tradicionalnog polaganja ispita, algoritamski model je procenjivao ocene učenika koristeći kombinaciju podataka o njihovom dosadašnjem učinku i rezultatima koje su postigli učenici u istim školama prethodnih godina. Ovaj pristup bio je osmišljen kako bi stvorio sistem koji bi bio ujednačen i koji bi sprečio nerealna poboljšanja ocena, jer su nastavnici sami predlagali ocene na osnovu procena o potencijalu učenika. Naime, vlada i Ofqual, regulatorno telo za ispite, želeli su da spreče inflaciju ocena i da se rezultati što vernije usklade s prethodnim godinama [25, 12].

Međutim, problem je nastao u načinu na koji je algoritam ponderisao podatke. Više pažnje se pridavalo istorijskim podacima o školi, a manje individualnom učinku učenika. Tako su učenici iz škola sa lošijim istorijskim rezultatima dobili niže ocene od onih iz škola sa tradicionalno boljim rezultatima, čak i ako su imali odlične ocene nastavnika. To je dovelo do toga da su mnogi učenici iz škola u siromašnijim delovima zemlje ili manjim zajednicama dobili znatno niže ocene nego što su očekivali, dok su učenici iz prestižnijih škola često zadržali visoke ocene. Tačnije, procene su pokazale da su deca iz socioekonomski ugroženih sredina bila znatno više pogođena algoritamskim sistemom ocenjivanja u poređenju sa decom iz bogatijih krajeva. Ono što je dodatno pogoršalo situaciju jeste činjenica da je algoritamski model odlučivanja favorizovao veće grupe učenika, dok je školama sa manjim brojem učenika dodeljivao nepredvidive rezultate. Individualne ocene učenika bile su praktično prepisivane iz prethodnih godina, čime su rezultati postali odraz kolektivnih performansi škole, a ne ličnog napretka učenika. To je u mnogim slučajevima značilo da su učenici u manjim razredima, poput onih u ruralnim ili siromašnijim školama, bili nepravedno kažnjeni, bez obzira na svoje akademske sposobnosti [25, 12, 30].

Javni pritisak je brzo rastao. Hiljade učenika i roditelja protestovali su protiv nepravednog algoritamskog model, a neki od njih su planirali pravne postupke. Kritike su stizale iz svih delova društva, uključujući političare, akademike i pravne eksperte, koji su isticali da algoritamski model odlučivanja nije uspeo da uzme u obzir specifične okolnosti pojedinačnih učenika i da je doveo do reprodukcije postojećih socijalnih nejednakosti. Na kraju, pod veoma jakim pritiskom, britanska vlada i regulator Ofqual povukli su algoritam i odlučili da se umesto toga koriste ocene koje su predložili nastavnici. Ova odluka usledila je nakon nekoliko dana velikih protesta i kritika, ali šteta je već bila učinjena: mnogi učenici su izgubili prilike za upis na željene univerzitete

zbog privremenog kašnjenja u ispravci ocena, a ceo događaj narušio je poverenje javnosti u digitalne sisteme ocenjivanja [25, 12].

Pregled literature

Pregled literature sadrži definiciju pravednosti koja se razvila u oblasti algoritamskog odlučivanja. Ove mere pravednosti su mahom preuzete iz modernih teorija pravde i preslikana u svoj matematički oblik [31]. Nakon ovog dela, biće opisani pristupi u rešavanju problema neželjene diskriminacije i nepravednosti u algoritamskom odlučivanju koji se javljaju u literaturi. Ovaj deo će se u velikoj meri osloniti na radove koji se bave pregledom oblasti [29, 31, 7] nepravednosti u algoritamskom odlučivanju, odnosno mašinskom učenju.

Mere pravednosti

Pravednost u algoritamskom odlučivanju je od suštinskog značaja za obezbeđivanje etičnosti i društvenog blagostanja u primeni algoritama mašinskog učenja [1]. Ono što bismo želeli da obezbedimo jeste da ovi sistemi ne oslikavaju pristrasnosti iz istorijskih podataka ili prave nove oblike diskriminacije. Da bismo mogli da uočimo nepravednost, moramo prvo da je izmerimo. U literaturi su se javili različiti pristupi za merenje i unapređenje pravednosti [29, 7].

Jedna od najčešće korišćenih mera pravednosti je demografska ravnopravnost ili različit uticaj (eng. *Disparate Impact*) [4, 61, 7]. Ova mera ispituje razliku u stopama dobijanja željenog resursa između povlašćenih i nepovlašćenih grupa. Grupe su definisane tzv. osetljivim atributima kao što su rasa, pol ili religija. Različit uticaj se računa kao odnos verovatnoća da pojedinac iz nepovlašćene grupe ostvari željeni ishod u poređenju sa pojedincem iz povlašćene grupe.

Odnosno:

$$DI = \frac{p(\hat{y} = 1|s = 1)}{p(\hat{y} = 1|s = 0)} \quad (1)$$

gde je $\hat{y}$ odluka modela algoritamskog odlučivanja koja može biti 1 (pojedinac je dobio željeni resurs) i 0 (pojedinac nije dobio željeni resurs), s predstavlja pripadnost nepovlašćenoj grupi (vrednost 1), odnosno povlašćenoj grupi (vrednost 0).

Idealna vrednost ove mere je 1, što ukazuje da obe grupe imaju jednak tretman u pogledu ishoda. Međutim, kada je ova vrednost manja od 1, to može ukazivati na pristrasnost u korist povlašćene grupe, dok vrednosti veće od 1 mogu ukazivati na pozitivnu diskriminaciju [24]. Na primer, ako algoritam za zapošljavanje daje prednost muškarcima u odnosu na žene, vrednost ove mere bila bi manja od 1, što bi ukazivalo na diskriminaciju prema ženama (kao što je pokazano u [11]).

Još jedna značajna mera pravednosti je statistička jednakost (eng. *Statistical Parity*), koja kvantifikuje apsolutnu razliku u proporcijama željenih ishoda između povlašćenih i nepovlašćenih grupa. Ova mera predstavlja modifikaciju različitog uticaja kada se racio pretvori u razliku, sa dodatkom apsolutne vrednosti razlike. Statistička jednakost predstavlja jednostavnu adaptaciju koja ima za cilj da obezbedi da krajnji ishodi ne favorizuju jednu grupu u odnosu na drugu.

$$SP = |p(\hat{y} = 1|s = 1) - p(\hat{y} = 1|s = 0)| \quad (2)$$

Kritika različitog uticaja i statističke jednakosti dolazi iz teorija pravde. Naime, iako uzmemo kao aksiom da su svi ljudi rođeni jednaki i da bi trebalo da imaju podjednake verovatnoće ostvarivanja željenog ishoda, zanemaruje se veliki broj faktora koji može da utiče na ishod, a posebno onih koji su objektivni kao što je sposobnost pojedinca i onih koji su rezultat srećnih okolnosti a na koje nije moguće uticati [52, 3]. Odnosno, pretpostavka jednakosti pojedinaca ima neželjene efekte trivijalnosti pravednog rešenja. Trivijalno matematičko rešenje jeste da se svakom pojedincu dodeli resurs ili da se svakom pojedincu ne dodeli resurs (uz pretpostavku da je tada definisana ova mera pravednosti). Ako bismo preslikali teorije pravde, ova mera bi predstavljala egalitarizam, koji je kritikovan kao demotivišuća za pojedince [6].

Jednake šanse (eng. *Equal Opportunity*) predstavlja jedan ključan koncept u pravednosti algoritamskog odlučivanja [4]. Ova mera obezbeđuje da pojedinci iz nepovlašćenih grupa imaju jednake šanse za postizanje željenih ishoda kao i pojedinci iz povlašćenih grupa. Na primer, u algoritmu koji odlučuje o kreditima, jednaka šansa bi podrazumevala da ljudi iz različitih rasa, pod jednakim uslovima, imaju istu verovatnoću da im kredit bude odobren. Ova mera je važna jer se ne fokusira samo na krajnje ishode, već i na to da li algoritam pruža jednake prilike za uspeh svim grupama, bez obzira na njihove osetljive atribute. Odnosno, ovo je prva mera koja uvodi koncept zasluge (eng. *merit*) za dobijanje željenog resursa.

$$EO = \frac{p(\hat{y} = 1|y = 1, s = 1)}{p(\hat{y} = 1|y = 1, s = 0)} \quad (3)$$

Nadovezujući se na koncept jednakih šansi, jednake verovatnoće (eng. *Equalized Odds*) zahtevaju da algoritam ne pravi razliku u verovatnoćama ishoda za bilo koju grupu, bez obzira na to da li je ishod pozitivan ili negativan. Ovo osigurava da algoritam ne diskriminiše ni u pozitivnim ($j = 1$) ni u negativnim odlukama ($j = 0$). U suštini, ova mera obezbeđuje da nijedna grupa nije u nepovoljnijem položaju u odnosu na bilo koji ishod. Ova mera se nakon velikih diskusija usvojila kao mera pravednosti u krivičnom pravosuđu u Sjedinjenim Američkim Državama [16], gde se osigurava da algoritam ne predviđa veću verovatnoću za ponovno izvršenje krivičnog dela kod pojedinaca iz jedne rase u odnosu na drugu.

$$EOdds = \frac{p(\hat{y} = 1|y = j, s = 1)}{p(\hat{y} = 1|y = j, s = 0)}, \forall j \in J \quad (4)$$

Uprkos brojnim merama pravednosti koje su razvijene kako bi se osigurala pravednost u algoritamskom odlučivanju, nemoguće je zadovoljiti sve ove mere istovremeno. To je zbog inherentnih sukoba između različitih definicija pravednosti. Na primer, različit uticaj i jednake šanse često dolaze u sukob jer različit uticaj zahteva jednaku stopu pozitivnih ishoda među grupama, dok jednake šanse zahtevaju da pojedinci iz različitih grupa, sa istim zaslugama, imaju jednake šanse za postizanje tih ishoda. I teorijski i praktično, zadovoljenje jedne mere može dovesti do narušavanja druge. Ovi konflikti pokazuju da ne postoji univerzalno rešenje za pravednost u algoritamskom odlučivanju i da je kompromis između različitih mera neizbežan [33, 1].

Pristupi u rešavanju nepravednosti u algoritamskom odlučivanju

U literaturi su se razvili različiti pristupi u rešavanju problema nepravednosti u algoritamskom odlučivanju [29, 7] koji se razlikuju u oblasti u kojima se vrši intervencija pravednosti. U procesu razvoja modela odlučivanja i u skladu sa pristrasnostima koje pojedinac, grupa ili društvo može da ima, intervencija može da se sprovede u fazi pripreme podataka (pretprocesiranje), u fazi modelovanja (prilagođavanje algoritama za pravednost) i nakon toga u fazi kalibracije verovatnoća ili prilagođavanja predviđanja (postprocesiranje) [7].

Početna priprema podataka (eng. data preprocessing) je prvi i najdirektniji korak u rešavanju nepravednosti u algoritamskom odlučivanju, jer se bavi uklanjanjem pristrasnostima pre nego što se model uopšte počne učiti. Ovaj

pristup se fokusira na manipulaciju ili modifikaciju podataka kako bi se smanjila pristrasnost u samom skupu podataka, jer su pristrasnosti koje se nalaze u istorijskim podacima često glavni izvor nepravednosti u algoritamskim sistemima.

Jedan od najjednostavnijih oblika pretprocesiranja je uklanjanje osetljivih atributa iz skupa podataka, kao što su pol, rasa ili religija, kako bi se sprečilo da algoritam koristi te informacije prilikom donošenja odluka [4]. Međutim, ovaj pristup ima svoje slabosti, jer čak i kada se osetljivi atributi uklone, algoritam može indirektno dedukovati te informacije iz drugih atributa. Na primer, atributi kao što su poštanski broj, obrazovanje ili zanimanje mogu delovati kao zamena za osetljive atribute i zadržati pristrasnosti prisutne u podacima.

Napredniji pristupi uključuju popravku podataka (eng. *data reparation*), gde se podaci aktivno ispravljaju da bi bili pravedniji. To može uključivati tehniku poznatu kao "masiranje" ciljne varijable (eng. *label massaging*) [21], gde se ishodi prilagođavaju tako da odražavaju pravedniju raspodelu. Na primer, ako je istorijski zapostavljana grupa u podacima nepravedno tretirana, neki podaci mogu biti izmenjeni kako bi osigurali da su ishodi pravedniji u odnosu na različite grupe. Alternativno, metode kao što je ponovno otežavanje (eng. *reweighting*) [22] mogu se koristiti kako bi se instancama iz nepovlašćenih grupa dodelila veća težina tokom treniranja modela, čime se smanjuje pristrasnost u samom procesu treniranja. Može se pristupiti i pravljenju takvih latentnih promenjivih da iz njih nije moguće identifikovati osetljivi atribut [62] ili promena vrednosti unutar jedne varijable tako da statistike prvog i drugog reda budu istovetne [17].

Pretprocesiranje ima tu prednost što se bavi problemom pristrasnosti na samom izvoru – u podacima. Međutim, postoji nekoliko izazova u ovom pristupu. Prvo, modifikacija podataka može ugroziti njihovu tačnost i verodostojnost, što može dovesti do lošijih performansi modela. Drugo, regulative poput GDPR-a postavljaju visoke zahteve za tačnost podataka, a manipulacija podacima može dovesti do pravnih problema. Ipak, za mnoge aplikacije gde su pristrasnosti duboko ukorenjene u istorijskim podacima, pretprocesiranje predstavlja važan prvi korak u pravljenju pravednijih algoritamskih sistema. [56, 7]

Prilagođavanje algoritama za pravednost predstavlja drugi pristup rešavanju nepravednosti u algoritamskom odlučivanju, koji se fokusira na sam proces treniranja modela. Za razliku od pretprocesiranja, gde se manipulacija vrši na podacima pre treniranja modela, prilagođavanje

algoritama za pravednost uključuje modifikaciju samih algoritama kako bi se postigli pravedniji ishodi tokom učenja. Ovaj pristup omogućava direktnu kontrolu nad balansiranjem tačnosti modela i pravednosti, što ga čini pristupom koji je najzastupljeniji u literaturi.

Jedan od najčešće korišćenih pristupa prilagođavanje algoritama za pravednost je dodavanje regularizacionih uslova koji penalizuju model ako uči nepravedne obrasce iz podataka. Na primer, algoritam može biti dizajniran tako da minimizira pristrasnost u predikcijama dok istovremeno optimizuje tačnost. Regularizacija pravednosti može koristiti različite metrike pravednosti, kao što su različit uticaj ili jednake šanse, i može biti dodata u ciljnu funkciju kao dodatni parametar koji algoritam mora optimizovati zajedno sa tačnošću. Jedan primer ovakvog pristupa je regularizator za uklanjanje predrasuda (eng. *Prejudice Remover Regularizer*) [23], koji se koristi za uklanjanje pristrasnosti prema osetljivim grupama tako što kažnjava model svaki put kada pristrasnost prema toj grupi bude detektovana.

Drugi pristup prilagođavanje algoritama za pravednost koristi se tehnikama suparničkog učenja (eng. *Adversarial Learning*) [60], gde se pored glavnog modela trenira i dodatni, suparnički model koji pokušava da predvidi osetljive atribute iz podataka. Cilj glavnog modela je da minimizira tačnost suparničkog modela, čime se smanjuje verovatnoća da će algoritam koristiti osetljive informacije u svojim predikcijama. Ovaj pristup omogućava postizanje visoke pravednosti bez eksplicitnog uklanjanja ili modifikovanja podataka, jer model postaje nesvestan osetljivih atributa tokom učenja.

Prilagođavanje algoritama za pravednost ima nekoliko ključnih prednosti. Prvo, omogućava optimizaciju pravednosti i tačnosti istovremeno, što može dovesti do boljih performansi modela u realnom svetu. Drugo, ovaj pristup je fleksibilniji jer omogućava algoritmu da nauči pravedne zakonitosti direktno iz podataka, bez potrebe za radikalnim promenama u podacima ili postprocesiranju predikcija. Međutim, prilagođavanje algoritama za pravednost može biti tehnički složenije za implementaciju, jer zahteva modifikacije u algoritmima za učenje, kao i pažljivo balansiranje između pravednosti i tačnosti. Takođe, zbog sukoba između mera pravednosti i tačnosti [33, 56, 1], može biti izazovno zadovoljiti kriterijume pravednosti i tačnosti istovremeno.

Naknadna obrada (postprocesiranje predviđanja) je pristup koji se primenjuje nakon što je model treniran i predikcije su napravljene. U ovom pristupu, model algoritamskog odlučivanja ostaje nepromenjen, ali se predikcije koje model generiše prilagođavaju kako bi bile pravednije.

Postprocesiranje je korisno kada je model već treniran, ali su potrebne korekcije u ishodima da bi se osigurala pravednost u donošenju odluka.

Jedan od najjednostavnijih načina za implementaciju postprocesiranja je modifikacija verovatnoća predikcija [55]. Postupci mogu biti jednostavni poput uprosečavanja ishoda između grupa, primene matematičkog modelovanja za određivanje optimalnog transporta kako bi se ostvarila jednakost raspodela verovatnoća, ali i druge.

Jedan popularan algoritam postprocesiranja je rekalibracija modela (eng. *Recalibration*) [59], gde se predikcije modela ponovo kalibrišu tako da zadovolje tražene kriterijume pravednosti. Na primer, ako model predviđa više pozitivnih ishoda za povlašćenu grupu u odnosu na nepovlašćenu grupu, postprocesiranje može prilagoditi predikcije tako da bude više pozitivnih ishoda za nepovlašćenu grupu. Jedan od glavnih izazova u postprocesiranju je kompromis između tačnosti i pravednosti. Dok postprocesiranje može efikasno ispraviti nepravedne ishode, može doći do smanjenja ukupne tačnosti modela jer se ispravljaju predikcije koje su inicijalno bile tačne. Takođe, postprocesiranje može biti nepraktično za sisteme u kojima se zahtevaju brze predikcije u realnom vremenu, jer zahteva dodatne korake nakon što je model doneo odluku [29, 7].

Razvijene pravedne metode algoritamskog odlučivanja

U narednim poglavljima biće ukratko prikazane razvijene metode algoritamskog odlučivanja. Prvo će biti prikazane pravedne višekriterijumske metode odlučivanja, a nakon toga metode koje koriste algoritme mašinskog učenja, a potpadaju u familije pretprocesiranja, prilagođavanja ili postprocesiranja.

Pravedne metode višekriterijumske analize odlučivanja

Jedan od osnovih zadataka odlučivanja jeste rangiranje. U skladu sa temom ovog rada, problem pravednosti u rangiranju analizira kako algoritmi mogu uticati na diskriminaciju određenih grupa ili pojedinaca kada su metode koje se prilagođavaju metode višekriterijumske analize odlučivanja [13].

U radu [40], autori se fokusiraju na prilagođavanje TOPSIS metode kako bi eliminisali nepoželjne diskriminacione efekte u donošenju odluka. Ovaj pristup uvodi novi koncept nazvan *fer težine* koji je osmišljen da minimizira uticaj kriterijuma koji stvaraju nepravedne ishode. Ova težina se koristi pri izračunavanju udaljenosti alternativa od idealnog i najgoreg rešenja.

Matematički, korekcija nepoželjnih dispariteta zasnovana je na linearnom matematičkom modelovanju.

Eksperimenti su izvedeni na sintetičkom i stvarnom skupu podataka (već opisanom COMPAS sistemu za podršku odlučivanju), koji obuhvata ocene recidivizma. Primenom predložene metode, za diskriminisanu grupu je došlo do prosečnog povećanja korisničke korisnosti za 0.0247, dok je za privilegovanu grupu došlo do smanjenja za 0.0688. Diskriminacija u procesu donošenja odluka smanjena je za više od 20%, dok je različiti uticaj porastao sa 0.7173 na 0.8587, što ukazuje na značajno smanjenje nepoželjne pristrasnosti. Odnosno, pokazalo se da manjim promenama u težinama kriterijuma možemo da dođemo pravednijih odluka.

Međutim, nije samo TOPSIS metoda unapređena. Pokazalo se da je moguće unaprediti i VIKOR metodu. U radu [14] istraženo je na koji način se težine kriterijuma mogu promeniti kako bi se postigla pravednija rangiranja. Pristup koji je odabran zahteva primenu kompleksnijih optimizacionih procedura. Razlog tome je činjenica da VIKOR metoda podrazumeva najmanje dva koraka do dolaska do rešenja, te su primenjene metaheuristike u cilju dolaska do pravednog rešenja. Odnosno, nakon početnog rangiranja, algoritam koristi metaheurističke algoritme, kao što su diferencijalna evolucija, optimizacija rojeva čestica i genetski algoritam, kako bi optimizovali težine kriterijuma. Cilj optimizacije je smanjenje razlike između rangiranja pre i posle optimizacije, koristeći različite merne funkcije, poput kvadratne i apsolutne razlike u rangovima, kao i minimizaciju nepodudaranja u parovima alternativa.

Eksperimenti su izvedeni na tri različita skupa podataka, uključujući skupove podataka o vozilima, zemljama Latinske Amerike i Nemačkim kreditima. Primena predložene metode značajno je smanjila disparitet u rangiranju između privilegovanih i diskriminisanih grupa, gde je različiti uticaj povećan iznad zakonskog minimuma od 0.8, dok je kompromis u tačnosti (u smislu promena u početnim rangovima) bio minimalan.

Kada se uporede rezultati sa drugim algoritmima, pravedna verzija VIKOR metode pokazuje najbolju ravnotežu između tačnosti i pravednosti, uz minimalne izmene u težinama kriterijuma. Treba da napomuneti da je cena koji pri tome plaćamo brzina izvršavanja algoritma. Odnosno, za navedeni pristup potrebno je zadati samo prag diskriminacije, a težine će se izmeniti da se dostigne željeni prag, dok druge metode zahtevaju veću prisutnost modelara odluke koji bi trebalo da isproba različite postavke dok se ne stigne do željenog rešenja.

Prelazak sa kardinalne na ordinalnu korisnost donosi novu dimenziju problema u metodama višekriterijumskog odlučivanja. Naime, korisnost alternative tada nosi veću subjektivnost, ali i veću šansu da korisnik ne iskaže svoje preferencije ispravno. Zbog toga se u radu [15] istražuje kako se uparene matrice poređenja alternativa u AHP metodi mogu prilagoditi da istovremeno zadovolje kriterijume pravičnosti i konzistentnosti. Nivo kompleksnosti se povećava sa činjenicom da je prostor mogućih vrednosti za svaku promenjivu diskretan. U radu je pokazano da je problem NP težak, te da je jedini prikladan način rešavanja ovog problema primenom genetskog algoritma koji za cilj da minimizira odstupanja između originalne i korigovane matrice procene, pri čemu se poštuju ograničenja konzistentnosti i pravičnosti.

Eksperimenti su izvedeni na sintetičkim podacima, gde su poređene matrice sa različitim vrednostima početne konzistentnosti i diskriminacije. U 99.5% slučajeva, algoritam je uspešno unapredio i konzistentnost i pravičnost matrica što ga čini robusnim za rešavanje navedenog problema.

Početna konzistentnost, izražena kroz odnos konzistentnosti (CR), ima značajan uticaj na složenost i uspeh optimizacije. Kada je početni CR visok (što znači niska konzistentnost), optimizacija postaje izazovnija. Odnosno, za veće matrice (broj alternativa veći od 6), bilo je teže postići željeni balans između konzistentnosti i pravičnosti. U situacijama sa visokom početnom konzistentnošću (nizak CR), algoritam je brže konvergirao i lakše postizao zadovoljavajući nivo pravednosti, jer nije bilo potrebe za velikim korekcijama u parnim poređenjima. Što je veća početna konzistentnost, to je manje promena bilo potrebno kako bi se postigli fer rezultati, što znači da je optimizacija bila stabilnija i rezultirala manjim odstupanjima od početnih preferencija. Eksperimentalni rezultati pokazuju da sa svakim povećanjem početne nekonzistentnosti, prosečna vrednost ciljne funkcije (odstupanje od inicijalne matrice) proporcionalno raste, što znači da je zahtev za korekcijama značajno veći u slučaju lošije početne konzistentnosti.

Sa druge strane početna nepravednost, izražena kroz indikator različiti uticaj, takođe igra ključnu ulogu u procesu optimizacije. Kada je početni različiti uticaj daleko ispod željene vrednosti od 0.8, algoritam mora značajno prilagođavati preferencijalne ocene kako bi eliminisao diskriminaciju. U slučajevima visoke početne nepravednosti (niske vrednosti različiti uticaj), bilo je teže postići balans između konzistentnosti i pravičnosti, jer je bilo potrebno napraviti značajne promene u vrednostima preferencijalnih poređenja kako bi se eliminisali diskriminatorni efekti. Autori su pokazali da,

čak i kada je početni različiti uticaj bio nizak (npr. 0.6), optimizacija je uspešno povećala pravednost na prihvatljiv nivo (oko 0.85). Međutim, veće korekcije u pravednosti obično su zahtevale veća odstupanja od početnih preferencija, što je ponekad dovelo do većih kompromisa u konzistentnosti.

Na kraju, predložena je i metoda FairAW [43] koja nadograđuje metodu jednostavnih aditivnih težina. Ključna komponenta FairAW metode je postprocesiranje težina kriterijuma sa ciljem minimizacije razlika između rezultata diskriminisanih i privilegovanih grupa, uz zadržavanje individualne pravednosti. U ovom slučaju individualna pravednost je definisana sa razlikom u težinama koja se ostvaruje promenom težina kriterijuma. Dakle, ideja je da se težine kriterijuma promene što manje, odnosno da se korisnost alternative promeni što manje, a da grupe ostvare što je moguće sličnije prosečne korisnosti.

Eksperimentalni rezultati pokazuju da FairAW metoda uspeva da poboljša grupnu pravednost (povećanjem različitog uticaja), dok održava relativno nizak nivo promena u individualnim ocenama, čime se balansiraju zahtevi za grupnu i individualnu pravednost. Pokazano je da početna nepravednost ima ključan uticaj na konačne rezultate. Kada je početni različiti uticaj veoma nizak, što ukazuje na značajnu razliku u ocenama između diskriminisanih i privilegovanih grupa, optimizacija mora izvršiti veće promene u težinama kriterijuma kako bi postigla zadovoljavajući nivo pravednosti. U takvim slučajevima, postoji veća verovatnoća da će se narušiti individualna pravednost, jer će promene u ocenama biti značajnije.

Napomenuli bismo i metode razvijene za kontrolisanje pravednosti u DEA metodi koje obezbeđuju jednakost različitog uticaja [44] i Rolsov koncept pravednosti [36] u DEA metodi. Odnosno, koncept pravednosti može da se ugradi i uspešno primeni i u drugim, srodnim disciplinama višekriterijumskog odlučivanja.

Pravedni modeli mašinskog učenja

Za razvoj pravednih modela mašinskog učenja bilo je potrebno ispitati kako i koji pristupi daju najbolji odnos između pravednosti i tačnosti. Za pristup pretprocesiranja podataka najpogodniji pristup deluje u otežavanju instanci u skupu podataka. Taj pristup je isproban u radovima [51, 49] gde su korišćene dve metode otežavanja: dodeljivanje težina podacima na osnovu njihovog potencijala da utiču na pravednost i metoda koja direktno uklanja nepravednost u podacima.

Eksperimenti su izvedeni na dva skupa podataka, tačnije, Adult i COMPAS. Rezultati su pokazali da su obe metode značajno povećale grupnu pravednost, dok je individualna pravednost imala varijabilne rezultate – u nekim slučajevima je poboljšana, a u nekim se pogoršala. Iako je došlo do blagog smanjenja tačnosti modela (do 3%), pravednost, merena kroz različiti uticaj, značajno je unapređena. Iako je pokazan kompromis između pravednosti i tačnosti, pokazano je i da je smanjenje tačnosti relativno malo u poređenju sa poboljšanjima u pravednosti, posebno na grupnom nivou. Drugim rečima, blago žrtvovanje tačnosti dovodi do zadovoljavajuće pravednih rešenja.

Umesto otežavanja instanci, može se pristupiti i putem dodatnog uzorkovanja [50. za poboljšanje pravednosti modela mašinskog učenja, uz minimalan uticaj na tačnost. Autori su primenili različite strategije dodatnog uzorkovanja, uključujući tradicionalno dodatnog uzorkovanja i SMOTE, na dva skupa podataka poznata po neželjenoj diskriminaciji. Fokus je bio na izjednačavanju ishoda za diskriminisane i privilegovane grupe kako bi se smanjila diskriminacija zasnovana na osetljivim atributima. Eksperimenti su izvedeni koristeći tri algoritma. To su Naivni Bajes, logistička regresija i algoritam nasumičnih stabala odlučivanja. Rezultati pokazuju da su tehnike dodatnog uzorkovanja povećale pravednost do 15%, dok je tačnost modela ostala stabilna, sa maksimalnim padom AUC-a od 3% i AUPRC-a od 10%. Najveće poboljšanje pravednosti postignuto je kada su izjednačeni željeni ishodi za obe grupe, dok su performanse algoritama uglavnom ostale na zadovoljavajućem nivou.

Iako su se ovi pristupi pokazali kao veoma uspešni, ono što im je nedostajalo jeste garancija uspeha u ostvarivanju pravednih odluka. Zbog toga se fokus usmerio na prilagođavanje algoritama za pravednost. Nadovezujući se na [61] razvijene su modifikacije logističke regresije osiguravaju da predikcije ne diskriminišu nesrazmerno određenu grupu ljudi u smislu izjednačenih šansi [42]. Za razliku od pristupa pretprocesiranja, kada su primenjene pravednosne restrikcije, tačnost modela (AUC) na Adult skupu podataka je opala sa 0.8977 na 0.8278, ali su mere pravednosti, poput različitog uticaja i jednakih mogućnosti, značajno poboljšane. Različiti uticaj se povećao sa 0.3932 na 0.9707, a metrike jednakih mogućnosti za željene i neželjene ishode takođe su se značajno poboljšale. Ovi rezultati su drastično bolji od pretprocesiranja podataka za pravednost i govore da je cena pravednosti veoma mala, te da ju je moguće dostići uz malo žrtvovanje tačnosti, što je potvrđeno i u drugim radovima [35]. Međutim, problem sa razvijenim

pristupima jeste što korisnik mora eksplicitno da zada hiper-parametar pravednosti koji nije nužno razumljiv za domenskog eksperta. Takođe, da bismo videli koliko smo zaista uspešni u pravednosti i tačnosti, moramo da isprobamo različite postavke eksperimenata što može biti naporno.

Međutim, napredniji algoritmi zahtevaju i naprednije optimizacione procedure. Tako je razvijen algoritam FAIR (eng. *Fair Adversarial Instance Re-weighting*) [32] koja kombinuje pristupe otežavanja i suparničkog učenja kako bi se poboljšala pravednost u modelima mašinskog učenja, naravno sa idejom minimalnog narušavanja tačnosti. Metoda koristi tri neuronske mreže: prva koja određuje težinu svake instance, druga koja predviđa vrednosti osetljivih atributa, a treća koja predviđa ciljne vrednosti. FAIR metoda dolazi u četiri varijante, pri čemu je osnovna razlika u načinu na koji se težine izračunavaju – kao skalarne vrednosti ili kao nasumične promenljive u probabilističkom okviru. Pokazano je da ovaj pristup dominira nad prethodno spomenutim pristupima zasnovanim na logističkoj regresiji u smisli tačnosti, ali i pravednosti. To nam sugeriše da pravednost nije linearan koncept, već da je veoma kompleksan [27] i da linearna granica odlučivanja nije najpogodnija za modelovanje mera pravednosti, bez obzira da li je reč o različitom uticaju, jednakosti šanse ili jednakih prilika. Metoda FAIR je donela još prednosti u odnosu na ostale metode koje su dostupne u literaturi. Jedna od najvažnijih inovacija FAIR metode je mogućnost interpretabilnosti modela. Koristeći težine instanci koje generišu neuronske mreže, modeli mogu da pruže uvid u to koje su instance pravedne, a koje nepravedne. Ovo je posebno korisno kada je potrebno doneti informisane odluke o tome koje instance iz skupa podataka treba isključiti kako bi se smanjila pristrasnost.

U cilju daljeg razumevanja odnosa između pravednosti i tačnosti, razvijeni su pristupi koji ispituju celokupan prostor optimalnih rešenja između pravednosti i tačnosti (tzv. formiraju Pareto front). Tačnije, u radu [26] Pareto front je ispitan kroz primenu NSGA-II (eng. *Nondominated Sorting Genetic Algorithm II*) koji služi za više-kriterijumsku optimizaciju. Cilj je bio da se istovremeno optimizuju dve metrike, od kojih je AUC korišćena kao mera tačnosti klasifikacionog modela, a različit uticaj kao mera pravednosti. NSGA-II algoritam koristi mehanizam sortiranja koji daje prednost nedominiranim rešenjima, tj. rešenjima koja su bolja ili jednaka u svim metrikama u odnosu na druga rešenja. Tokom eksperimenata, algoritam NSGA-II generisao je veliki broj rešenja sa različitim kombinacijama tačnosti i pravednosti, te je pokazano da ovaj pristup može da kreira dominantna rešenja u poređenju sa modelima fer logističke regresije opisanih ranije [61, 42, 35]. NSGA-II je

uspeo da postigne skoro savršenu pravednost (DI ~1) sa AUC blizu 0.87, dok su pristupi zasnovani na fer logističkoj regresiji postizali veću tačnost (AUC~0.9), ali uz značajan gubitak u pravednosti.

Međutim, postavlja se pitanje da li su mere pravednosti u skladu sa konceptom pravednosti koji se može pronaći u društvenim nauka, poput literature iz etike, filozofije, prava i slično. U radu [20] govori se kako uobičajene matematičke mere pravednosti, poput različitog uticaja ili jednakih šansi, koriste zbog njihove jednostavnosti i lakoće implementacije, ali te mere često predstavljaju pojednostavljenu aproksimacije stvarnog koncepta pravednosti. Zaključak je da ključna tačka neusklađenosti leži u tome što ove matematičke mere ne uspevaju da obuhvate složenost društvenih problema i njihove implikacije na određene grupe ljudi i da bi trebalo da se razviju drugačiji pristupi, pristupi koji dolaze iz društvene nauke i koje nude dublje i šire razumevanje pravednosti koje uključuje različite oblike socijalne nejednakosti, istorijske faktore i strukturalne barijere, što matematički modeli često zanemaruju. Rad naglašava potrebu da se metrike pravednosti u mašinskom učenju usklade sa ovim širim socijalnim konceptima, umesto da se oslanjaju na matematički pogodna, ali ograničena rešenja. To bi moglo uključivati upotrebu metoda kao što su učenje sa pojačavanjem (eng. *Reinforcement Learning*), formalizacija pojma pravednosti bez zavisti (eng. *Envy-Free Fairness*) ili proceduralne pravednosti, koji bi bolje reflektovali ciljeve društvene pravednosti.

Kao dodatnu motivaciju za prethodni rad, sprovedeni su eksperimenti koji ispituju da li aproksimacije mera pravednosti zapravo ostvaruju pravu pravednost kako je ona definisana. U radovima [37, 46] se detaljno analizira neusklađenost između matematičkih mera pravednosti i stvarnog koncepta pravednosti koji bi trebalo da bude postignut u algoritmima mašinskog učenja. Glavni problem koji se ističe u radovima dolazi iz činjenice da linearna aproksimacija nije precizna, te, posledično, rezultuje u pogrešnim zaključcima o pravednosti modela.

Na osnovu eksperimentalnih rezultata prikazanih u radu, primećeno je da postoji značajan jaz između aproksimiranih i stvarnih vrednosti pravednosti. Na primer, za COMPAS skup podataka, jaz između aproksimiranog i stvarnog različitog uticaja može da iznosi 21%, dok je za Adult skup podataka taj jaz još veći i dostiže čak 35%. Ono što još više šteti zaključcima jeste da su aproksimacije uvek optimističnije, te se donosilac odluke može zavarati i pomisliti da je veoma fer u donošenju odluka, a da zapravo nije.

Upravo iz poslednje navedenog razloga pravednost je ispitana na drugačiji način, odnosno, kroz odsustvo zavisti. U radu [39], zavist postoji kada bi pojedinac bio zadovoljniji ishodom koji je namenjen pojedincu iz druge grupe. Da bi model eliminisao zavist, uvedena je penalizaciona funkcija za zavist u model logističke regresije, čime je modifikovana standardna funkciju gubitka logističke regresije. Ova penalizaciona funkcija je linearno aproksimirana zbog nelinearnosti zavisti i dodata u funkciju cilja., te je uveden kontrolni hiperparametar koji omogućava da se reguliše nivo zavisti unutar modela. Eksperimenti su pokazali da predloženi model sa odsustvom zavisti postiže bolje rezultate u smislu pravednosti u poređenju sa standardnom logističkom regresijom, kao i posebnim logističkim regresijama po grupama. Pokazalo se da se ovim pristupom dobija i veća tačnost i bolja pravednost. Međutim, računski je veoma kompleksna i potrebno je dodatno unapređenje kako bi se ovakvi modeli dalje primenjivali.

Pored navedenih pristupa izdvajaju se i [38] koji modeluje pravednost kroz teoriju informacije, kao i [41] koji rešava problem pravednosti kroz lanac klasifikatora u problemima kada postoji veći broj izlaznih promenjivih.

Na kraju, pokušaj da se koriguju predikcije je takođe sproveden u radu [45]. Tačnije sprovedene su tri tehnike postprocesiranja predikcija za poboljšanje pravednosti algoritma logističke regresije. Koristile su se tehnike ujednačavanja šansi, kalibrisanja verovatnoća i šansi po grupama, kao i opcija odbijanja odluka. Rezultati su pokazali da primena navedenih tehnika značajno poboljšava pravednost modela. Tačnije, različiti uticaj je povećana za oko 40%, dok je jednakost prilike poboljšana za približno 60%. Međutim, cena greške je značajno veća u poređenju i sa metodama pretprocesiranja podatka i sa pristupom prilagođavanja algoritama za pravednost.

Zaključak

Algoritamska pravednost predstavlja složen i višedimenzionalan problem koji dobija sve veći značaj u eri mašinskog učenja i automatizacije donošenja odluka. Iako algoritmi mogu ubrzati i unaprediti mnoge procese, oni mogu istovremeno i nesvesno reprodukovati društvene nepravde i pristrasnosti koje su već prisutne u podacima na kojima su trenirani. Problemi poput onih u Amazonovom algoritmu za zapošljavanje ili sistemu COMPAS za procenu rizika ukazuju na ozbiljne posledice koje nepravedni algoritmi mogu imati na živote ljudi, naročito pripadnika marginalizovanih grupa [1, 4].

U radu su razmotrene ključne mere pravednosti i izazovi u njihovoj primeni. Takođe, diskutovani su pristupi rešavanju problema algoritamske

pravednosti. Svaka od ovih tehnika ima svoje prednosti i slabosti, a često zahteva kompromis između tačnosti i pravednosti [7].

Iako ne postoji univerzalno rešenje koje bi moglo zadovoljiti sve definicije pravednosti [33], razvijeni su pristupi koji omogućavaju postizanje balansiranih rešenja. Prilagođavanje algoritama za pravednost, posebno kroz upotrebu tehnika suparničkog učenja i regularizacije, pokazalo se kao jedan od efikasnijih pristupa za smanjenje pristrasnosti, dok metode naknadne obrade nude fleksibilnost u prilagođavanju predikcija nakon što su modeli trenirani. Ovaj rad prikazuje razvijene pristupe i sumarizuje zaključke nastale iz niza istraživanja u ovoj oblasti.

Međutim, izazovi ostaju, kako u definisanju i merenju pravednosti, tako i u implementaciji pravednih modela u praksi. Budući radovi treba da se fokusiraju na integraciju šireg spektra društvenih i etičkih koncepata pravednosti u matematičke modele, kao i na razvoj tehnika koje omogućavaju transparentnost i odgovornost u automatizovanom donošenju odluka.

Reference

1. Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020, January). Roles for computing in social change. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 252-260).
2. Amartya, S. (2009). *The idea of justice* (p. 253). London: Penguin Books.
3. Amartya, S. (2017). What do we want from a theory of justice?. In *Theories of Justice* (pp. 27-50). Routledge.
4. Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.
5. Bhopal, K., & Myers, M. (2020). The impact of COVID-19 on A Level students in England. *SocArXiv*.
6. Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency* (pp. 149-159). PMLR.
7. Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), 1-38.
8. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153-163.
9. Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, *63*(5), 82-89.
10. Coleman, J. S., Frankel, B., & Phillips, D. L. (1976). Robert Nozick's anarchy, state, and utopia. *Theory and Society*, *3*(3), 437-458.

11. Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296-299). Auerbach Publications.
12. Denes, G. (2023). A case study of using AI for General Certificate of Secondary Education (GCSE) grade prediction in a selective independent school in England. *Computers and Education: Artificial Intelligence*, *4*, 100129.
13. Dodevska, Z., Delibašić, B. & Radovanović, S. (2020) Decision making with Fair Ranking. In *Proceedings of the XVII International Symposium SymOrg* (pp. 278-282), Zlatibor, Serbia, September 7-10, Serbia.
14. Dodevska, Z., Petrović, A., Radovanović, S., Delibašić, B. (2022) Changing criteria weights to achieve fair VIKOR ranking: a postprocessing reranking approach. *Autonomous Agents and Multi-Agent Systems, 37*(9), pp. 1-44.
15. Dodevska, Z., Radovanović, S., Petrović, A., & Delibašić, B. (2023). When Fairness Meets Consistency in AHP Pairwise Comparisons. *Mathematics, 11*(3), 604.
16. Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.
17. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259-268).
18. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50-57.
19. Hagendorff, T. (2022). Blind spots in AI ethics. *AI and Ethics*, *2*(4), 851-867.
20. Jovanović, M., Radovanović, S., Delibašić, B. (2022) Misalignment of Fairness in Machine Learning. In Proceedings of International Scientific Conference Emerge 2022 (pp. 10-11). December 16-18, Belgrade, Serbia.
21. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, *33*(1), 1-33.
22. Kamiran, F., Karim, A., & Zhang, X. (2012, December). Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining* (pp. 924-929). IEEE.
23. Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23* (pp. 35-50). Springer Berlin Heidelberg.

24. Kannan, S., Roth, A., & Ziani, J. (2019, January). Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 240-248).
25. Kelly, A. (2021). A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal*, *47*(3), 725-741.
26. Kovačević, A., Vukićević, M., Radovanović, S., Suknović, M. & Delibašić, B. (2021, September) Fair and Accurate Logistic Regression with Multiobjective Metaheuristic optimization. In *Proceedings of XLVIII Symposium of Operational Research - SYM-OP-IS 2021* (pp. 671-676), September 20-23, Banja Koviljača, Serbia.
27. Lohaus, M., Perrot, M., & Von Luxburg, U. (2020, November). Too relaxed to be fair. In *International Conference on Machine Learning* (pp. 6360-6369). PMLR.
28. Malek, M. A. (2022). Criminal courts' artificial intelligence: the way it reinforces bias and discrimination. *AI and Ethics*, *2*(1), 233-245.
29. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1-35.
30. Murphy, R., & Wyness, G. (2020). Minority Report: the impact of predicted grades on university admissions of disadvantaged groups. *Education Economics*, *28*(4), 333-350.
31. Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, *55*(3), 1-44.
32. Petrović, A., Nikolić, M., Radovanović, S., Delibašić, B., & Jovanović, M. (2022). FAIR: Fair adversarial instance re-weighting. *Neurocomputing*, *476*, 14-37.
33. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, *30*.
34. Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, *32*, 6363-6381.
35. Radovanović, S., & Ivić, M. (2023). Enabling Equal Opportunity in Logistic Regression Algorithm. *Management: Journal of Sustainable Business and Management Solutions in Emerging Economies*, *28*(2), 55-66.
36. Radovanović, S., Delibašić, B., Marković, A., & Suknović, M. (2022). Achieving MAX-MIN Fair Cross-efficiency scores in Data Envelopment Analysis.
37. Radovanović, S., Delibašić, B., Suknović, M. (2022) Do we Reach Desired Disparate Impact with In-Processing Fairness Techniques? *Procedia Computer Science 214*, 257-264.
38. Radovanović, S., Delibašić, B., Suknović, M. (2022) Empirical analysis of Information Theory point of view on Equal Odds Fairness Measure. In *Proceedings of 8th International Conference on Decision Support System*

Technology – ICDSST 2022 (pp. 43-49). May 23-25, Thessaloniki, Greece.
39. Radovanović, S., Delibašić, B., Suknović, M. (2022) Envy-free Fair Student Dropout Prediction. In *Proceedings of XVIII International Symposium of Organizational Sciences - SymOrg 2022* (pp. 42-43). June 11-14, Belgrade, Serbia.
40. Radovanović, S., Petrović, A., Delibašić, B. & Suknović, M. (2021, October) Eliminating Disparate Impact in MCDM: The case of TOPSIS. In *Proceedings of Central European Conference on Information and Intelligence Systems - CECIIS 2021* (pp. 275-282). October 13-15, Varaždin, Croatia.
41. Radovanović, S., Petrović, A., Delibašić, B., & Suknović, M. (2023). A fair classifier chain for multi-label bank marketing strategy classification. *International Transactions in Operational Research*, *30*(3), 1320-1339.
42. Radovanović, S., Petrović, A., Delibašić, B., Suknović, M. (2020). Enforcing fairness in logistic regression algorithm. In *Proceedings of 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, August 24-26, 2020. Novi Sad, Serbia, pp. 1-7, DOI: 10.1109/INISTA49547.2020.9194676.
43. Radovanović, S., Petrović, A., Dodevska, Z., & Delibašić, B. (2023). FairAW–Additive weighting without discrimination. *Intelligent Data Analysis*, *vol. 27*(4), pp. 1023-1045.
44. Radovanović, S., Savić, G., Delibašić, B., & Suknović, M. (2022). FairDEA—Removing disparate impact from efficiency scores. *European Journal of Operational Research*, *301*(3), 1088-1098.
45. Radovanović, S., Suknović, M., Ivić, M. (2020) Can we adjust predictions to be Fair? Postprocessing of the prediction model. In *Proceedings of the XVII International Symposium SymOrg* (pp. 263-269), Zlatibor, Serbia, September 7-10, Serbia.
46. Radovanović, S., Vukićević, M., Suknović, M. (2024) Are we really addressing fairness in machine learning algorithms? In *Proceedings of the XIX International Symposium of Organizational Sciences – SymOrg 2024*, pp. 291-296. Zlatibor, Serbia.
47. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020, January). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 469-481).
48. Raji, I. D., & Buolamwini, J. (2022). Actionable auditing revisited: Investigating the impact of publicly naming biased performance results of commercial AI products. *Communications of the ACM*, *66*(1), 101-108.
49. Rančić, S., Radovanović, S., & Delibašić, B. (2021, May). Improving fairness in machine learning models with instance weighting. In Proceedings of the 7th International Conference on Decision Support

System Technology - ICDSST 2021 (pp. 61). May 26-28, Loughborough University, Loughborough, UK.
50. Rančić, S., Radovanović, S., & Delibašić, B. (2021, May). Investigating Oversampling Techniques for Fair Machine Learning Models. In International Conference on Decision Support System Technology (pp. 110-123). Springer, Cham.
51. Rančić, S., Radovanović, S., Suknović, M. (2020) Effects of Data Preprocessing for Fairness in Machine Learning. In *Proceedings of the XVII International Symposium SymOrg* (pp. 255-262), Zlatibor, Serbia, September 7-10, Serbia.
52. Rawls, J. (1971). A theory of justice. *Cambridge (Mass.)*.
53. Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170362.
54. Stigilitz, J. E. (2012). The price of inequality.
55. Tifrea, A., Lahoti, P., Packer, B., Halpern, Y., Beirami, A., & Prost, F. FRAPPÉ: A Group Fairness Framework for Post-Processing Everything. In *Forty-first International Conference on Machine Learning*.
56. Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567.
57. Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ*, *17*, 131.
58. Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, *96*(4), 114-123.
59. Xian, R., Yin, L., & Zhao, H. (2023, July). Fair and optimal classification via post-processing. In *International Conference on Machine Learning* (pp. 37977-38012). PMLR.
60. Xu, H., Liu, X., Li, Y., Jain, A., & Tang, J. (2021, July). To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning* (pp. 11492-11501). PMLR.
61. Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, April). Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics* (pp. 962-970). PMLR.
62. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In *International conference on Machine Learning* (pp. 325-333). PMLR.

Sandro Radovanović*

University of Belgrade – Faculty of Organizational Sciences, Serbia

Andrija Petrović

University of Belgrade – Faculty of Organizational Sciences, Serbia

Boris Delibašić

University of Belgrade – Faculty of Organizational Sciences, Serbia

Milija Suknović

University of Belgrade – Faculty of Organizational Sciences, Serbia

*sandro.radovanovic@fon.bg.ac.rs

FAIRNESS AND EQUALITY IN ALGORITHMIC DECISION MAKING

Abstract The paper explores the issue of algorithmic fairness, particularly in the context of automated decision-making systems and machine learning algorithms, which are increasingly used in areas such as employment, justice, and education. Through the analysis of well-known examples of unfair algorithms, it demonstrates how biased data and models can lead to discriminatory decisions. The paper also includes a literature review on addressing algorithmic unfairness, covering approaches such as data preprocessing, algorithm adjustment, and prediction post-processing. It presents the key results achieved by the group of authors in this field. Experimental results show that there is an inevitable trade-off between fairness and accuracy, regardless of the approach used, but it has been shown that a slight reduction in accuracy can significantly improve fairness, particularly at the group level. However, there is a need for further research into methods that combine mathematical and social aspects of fairness, as well as the development of algorithmic systems that ensure transparency and accountability in decision-making.

Key words: *Algorithmic fairness, algorithmic decision-making, machine learning, data preprocessing, algorithm adjustment, post-processing*