UDC: 004.85 Info M str. 28-34

ПРИМЕНА ТЕХНИКА ЗА ПОСТИЗАЊЕ ПРАВЕДНИХ МОДЕЛА МАШИНСКОГ УЧЕЊА
APPLICATION OF TECHNIQUES FOR ACHIEVING FAIRNESS IN MACHINE LEARNING MODELS

Милица Перишић, Сандро Радовановић, др Борис Делибашић

РЕЗИМЕ: Модели машинског учења постају све присутнији у људској свакодневници и широко се примењују у различитим индустријама. Одлуке се све чешће препуштају моделима који уче из доступних података о појави која се испитује. Подаци описују људско понашање које је кроз историју често имало дискриминишући карактер према одређеним групама друштва. Данас смо сведоци покушаја да се уклони дискриминација где год је то могуће, али како модели машинског учења уче на основу историјских података дешава се да се дискриминација из података научи и примењује. Због тога постаје јако битно направити праведан модел. Међутим, свакако да није довољно да модел буде само праведан. Основни циљ остаје да модел буде тачан. Изазов лежи у проналажењу компромиса између тачности и праведности модела. Овај рад ће се фокусирати на давање одговора на питање колика је цена повећања праведности модела машинског учења. Рад ће дати преглед тренутно најважнијих техника за постизање фер модела машинског учења, начина евалуације као и њихових цена како би се објаснио компромис између тачности и праведности модела.

КЉУЧНЕ РЕЧИ: машинско учење, дискриминација, праведност алгоритама машинског учења

ABSTRACT: Machine learning models are becoming more and more present in everyday life and are widely applied in various industries. Decisions are increasingly left to models which learn from available data on the phenomenon being examined. The data describe human behavior that throughout history has often had a discriminatory character towards certain groups in society. Today, we are witnessing attempts to eliminate discrimination wherever possible, but as machine learning models learn from historical data, discrimination from data is learned and applied. Therefore, it becomes very important to make a fair model. However, it is certainly not enough for the model to be just fair. The main goal remains to make the model accurate. The challenge lies in finding a compromise between the accuracy and fairness of the model. This paper will focus on answering the question of what the cost of increasing the fairness of a machine learning model is. The paper will provide an overview of the currently most important techniques for achieving fair machine learning models, evaluation methods as well as costs to explain the trade-off between model accuracy and fairness.

KEY WORDS: Machine learning, discrimination, fair machine learning

1. УВОД

Алгоритми машинског учења су данас готово свеприсутни. Користе се у пословним применама за предвиђање одлазака клијената или враћања кредита, али се користе и као помоћ свакодневном животу у паметним уређајима. Предности које алгоритми машинског учења доводе су пре свега брзина одлуке која се драстично смањује, али и тачност одлуке. Ове предности се добијају зато што су алгоритми машинског учења у стању да посматрају и обрађују већи број особина проблема који се посматра [23]. Начин на који алгоритми машинског учења раде се заснива на обележеним историјским подацима. Односно, алгоритми машинског учења покушавају да науче правила која су се јављала у историјским подацима која су доводила до жељеног исхода. У циљу генерализације научених правила дефинише се функција грешке за коју желимо да буде што нижа [5]. Међутим, модели одлучивања могу систематски да награђују једну групу људи зато што су културолошки или историјски били награђивани. Додатни проблем је што та група људи поседује изражену особину по којој не би требало одлука да се доноси, као што су пол, раса или вероисповест. У том случају, одлуке које доносе модели машинског учења могу бити неетичке, те резултовати правним последицама [1].

Решавање проблема нежељене дискриминације је вишеструко изазован. Први и најбитнији проблем јесте откривање где је дискриминација настала. Дискриминација се може појавити као резултат културолошких појава у друштву и као такви да се огледају у историјским подацима. На пример, студије инжењерства чешће уписује мушки пол зато што је културолошки уверење да је инжењер особа мушког пола [19]. Због тога ће се и у подацима осликати већи број кандидата мушког пола у односу на женски пол. Затим, узрок дискриминације може бити процес прикупљања података. Наиме, постоје примери да се подаци свесно или несвесно занемарују приликом анализе [28]. Пример је *COMPAS*, софтвер за подршку одлучивању у судству. Задатак система је да препоручи казну на основу особина појединца и претходних кривичних дела. Модел је научен на скупу података који садржи знатно мање припадника беле расе у односу на припаднике негроидне расе. Разлог није што бела раса чини мање злочина, већ што полиција није бележили све злочине беле расе [7]. На крају, извор дискриминације може бити и сам доносилац одлуке током одређивања исхода. На пример, приликом запошљавања доносилац одлуке може фаворизовати припаднике мушког пола, те ће алгоритам машинског учења научити ту дискриминацију [25]. Проблем је додатно изазован јер није довољно само направити праведан модел. Потребно је да он и даље задовољава основни циљ – да буде тачан [4].

Циљ овог рада јесте приказивање приступа у решавању нежељене дискриминације у моделима и алгоритмима машинског учења. У литератури се могу пронаћи технике којима се смањује дискриминација у моделима машинског учења. Технике су подељене у три групе – технике претпроцесирања, постпроцесирања и технике при-

лагођавања алгоритама машинског учења [7]. У овом раду испитаће се како примена наведених техника смањује могућност нежељене дискриминације. Битно је напоменути да свака техника даје различите резултате у зависности од скупа података који се обрађује и модела машинског учења који се гради. Такође, примена техника смањује тачност модела и потребно је наћи компромис између ова два циља. Кроз рад биће коришћена библиотека *AIF360* [2] која представља скуп алата отвореног кода за откривање, разумевање и ублажавање алгоритамске пристрасности. Крајњи резултат рада биће преглед цена коришћења различитих техника за повећање праведности модела машинског учења (енгл. *Cost of fairness*), што ће објаснити горе поменути компромис.

Остатак рада је структуриран на следећи начин. Преглед литературе је приказан у другом поглављу. Затим је приказана методологија рада, односно кратак опис података и поставка експеримената. Резултати и дискусија резултата у поглављу четири. На крају, приказан је закључак.

2. ПРЕГЛЕД ЛИТЕРАТУРЕ

У последњих 10 година све више је научних радова који се баве овом темом. Смањење дискриминације у друштву, па тако и у моделима машинског учења је постало значајан предмет изучавања. Да бисмо разумели начине како се исправља нежељена дискриминација, потребно је опишемо шта је праведност и како се она може измерити.

Први поглед на праведност је тзв. појединачна правда (енгл. *Individual fairness*) [22]. За модел кажемо да је праведан ако за произвољну меру удаљености и за веома сличне инстанце скупа података и према мери удаљености (односно) важи да је исход модела машинског учења веома сличан, односно . Алгоритми машинског учења (или модели алгоритамског доношења одлука у општем случају) имају проблем са појединачном правдом само у случају високе варијансе модела. Уколико се модел машинског учења правилно регуларизује, не постоји могућност да исход модела значајно разликује за сличне инстанце у скупу података [16].

Имајући то у виду, алгоритми машинског учења се чешће баве тзв. групном правдом (енгл. *Group fairness*) [8]. Групна правда или праведност представља ситуацију у којој не желимо да дозволимо да лична особина појединца која не може, или ју је тешко променити, утиче на исход модела. Те особине су нпр. пол, раса или вероисповест. Тачније, одлука модела машинског учења би требало да буде независна од личних особина. Уколико дефинишемо две групе, дискриминисану () и привилеговану () онда можемо измерити нежељену дискриминацију користећи различит утицај (формула 1) или разлику статистичких паритета [10] (формула 2):

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

$$AVG_o = \frac{\left(FPR_d - FPR_p\right) + \left(TPR_d - TPR_p\right)}{2} \quad (2)$$

где је математичко очекивање, предвиђен исход, а припадност групи. Различит утицај представља однос очекиваног жељеног исхода модела машинског учења, док разлика статистичких паритета представља разлику очекиваног жељеног исхода модела машинског учења. Идеална вредност различитог утицаја је један, док је разлика статистичког паритета једнака 0. Све друге вредности, било веће или ниже представљају постојање дискриминације. Дозвољене вредности за су између 0,8 и 1,25, док су дозвољене вредности за између -0,2 и 0,2 [15]. Поред ових мера, могу се користити и друге мере попут једнакост разлика у приликама (формула 3) и просечна разлика у приликама (формула 4):

$$EO = TPR_d - TPR_p \quad (3)$$

$$AVG_o = \frac{\left(FPR_d - FPR_p\right) + \left(TPR_d - TPR_p\right)}{2} \quad (4)$$

где је одзив дискриминисане групе, одзив привилеговане групе, проценат лажних упозорења дискриминисане групе и проценат лажних упозорења привилеговане групе. Једнаке разлике у приликама представљају разлику стопе тачно предвиђене позитивне класе између дискриминисаних и привилегованих група [10]. Идеална вредност је 0. Вредност подразумева већу корист за привилеговану групу, а вредност подразумева већу корист за дискриминисану групу. За модел се каже да је праведан ако је вредност између -0,1 и 0,1. Просечна разлика у приликама представља просечну разлику стопа погрешно предвиђене позитивне класе и стопе тачно предвиђене позитиве класе између дискриминисане и привилеговане група. Идентично тумачење важи и за [2].

Са наведеним мерама дискриминације могуће је измерити праведност модела пре и после његове модификације. Постизање праведних модела машинског учења подразумева примену неке од техника којом се елиминише пристрасност у подацима. Развијена су три приступа.

Први приступ подразумева да се прилагоде подаци пре процеса учења модела, тј. ***претпроцесирање података***. Један од најпознатијих приступа је **уклањање различитог утицаја** (енгл. *Disparate Impact Remover*) [9]. Ова техника претпроцесирања мења иницијални скуп података тако да није могуће разликовати дискриминисану и привилеговану групу у подацима за сваки посматрани атрибут. Ова корекција података се назива геометријска корекција зато што након прилагођавања податак појединац неће бити у стању да посматрањем инстанци разликује дискриминисану и привилеговану групу. Корекција се постиже померањем расподеле података обе групе ка аритметичкој средини целог скупа података. Формалније, вектор вредности атрибута се пресликава у тако да важи једначина 5:

$$\hat{C} = F_A^{-1}(F_S(C)) \quad (5)$$

где је нова расподела коју захтевамо (она која представља атрибут без знања о осетљивом атрибуту), инверзна функција расподеле и функција расподеле атрибута за задати атрибут припадности групи . Након оптимизације вредности инстанци које припадају привилегованој групи се смањују, док се вредности инстанци које припадају дискриминисаној групи повећавају. Фелдман и остали [9] додају параметар ниво поправке (*енгл. Repair level*), који узима вредност од 0 до 1. Нула представља случај у којем се вредности атрибута не мењају, док је 1 потпуна поправка и расподеле атрибута за различите групе се поклапају те је немогуће направити разлику на основу вредности истих.

Друга техника претпроцесирања податак која се издваја јесте праведно **отежавање инстанци** (***енгл.*** *Reweighing*) [12]. Ова метода не мења почетни скуп података, већ инстанцама додаје тежине. Односно, дискриминисаним инстанцама би требало доделити већу важност него привилегованим. Модел у фази учења узима у обзир израчунате тежине чиме исправља дискриминаторно понашање модела. Нека је скуп података на којим се учи модел, сензитивни, односно заштићени атрибут и класа која се предвиђа. Уколико су и статистички независни, тада је очекивана вероватноћа . Међутим, ако се две вероватноће разликују, онда то указује на постојање пристрасности ка једној од класа. Тада коригујемо важност инстанце. За то користимо формуле 6, 7 и 8.

$$P_{exp}(s = b \text{ and } Class = 1) := \frac{|\{X \in D \mid X(s) = b\}|}{|D|} \times \frac{|\{X \in D \mid X(Class) = 1\}|}{|D|} \quad (6)$$

$$P_{obs}(s = b \text{ and } Class = 1) := \frac{|\{X \in D \mid X(s) = b \text{ and } X(Class) = 1\}|}{|D|} \quad (7)$$

$$W(X) := \frac{P_{exp}(s = b \text{ and } Class = 1)}{P_{obs}(s = b \text{ and } Class = 1)} \quad (8)$$

Тежине се на овај начин додељују свим комбинацијама сензитивног атрибута и циљног атрибута. Нови скуп података, којем су дате тежине, не указује на пристрасност према некој од група јер је она уклоњена увођењем тежина инстанци. На тај начин се избегава случај у којем модел учи да дискриминише неку од група.

Други приступ јесте ***прилагођавање алгоритама машинског учења***. Ова приступ модификује алгоритам машинског учења тако да не постиже дискриминаторно понашање. Ово се може постићи регуларизацијом функције циља [14]. Регуларизација представља технику којом се контролише комплексност модела машинског учења. Тачније, врши се пенализација тачности модела уколико она доводи до дискриминаторног понашања. Други приступ јесте додавањем ограничења у математички модел [29, 21]. На крају, постоји приступ супарничког учења [30]. Заинтересовани читаоци се упућују на [18].

На крају, могуће је ***прилагодити излаз из модела*** машинског учења тако да он буде праведан. У том случају се поступак учења модела машинског учења спроводи на идентичан начин као и у класичним применама са изузетком да се након учења модела примењује техника која прилагођава било вероватноћу догађаја, било исход тако да се постиже праведнија одлука. Први приступ јесте **класификација заснована на опцији одбацивања** (енгл. *Reject Option Classification*) [12]. Ова техника подразумева нешто другачији приступ, од чистог мењања прага класификације. Уводи се критична зона око прага класификације у којој инстанце које припадају дискриминисаној и привилегованој групи буду класификоване у позитивну и негативну класу, тим редоследом. Тај исход се додељује без обзира да ли инстанаца из привилеговане групе има већу вероватноћу исхода од инстанце из дискриминисане групе. Односно, да би се смањила дискриминација, оне опсервације које се налазе у критичној зони су класификоване на следећи начин: уколико опсервација припада дискриминисаној групи додаје јој се позитивна класа, а уколико припада привилегованој групи додаје јој се негативна класа. Опсервације ван критичне зоне су класификоване према стандардном начину доношења одлуке [13].

Поред наведене методе издваја се и метода изједначавања прилике (енгл. *Equalized Odds Postprocessing*) [10]. Циљ методе јесте постизање једнакости разлике у приликама између привилеговане и дискриминисане групе за већ научени бинарни класификатор. Не мења се процес тренирања класификатора, већ се из постојеће предвиђене класе или вероватноћа припадности класама прави недискриминаторни модел. Међутим, потребно је да модел задовољава и одређени ниво тачности, те се уводи функција минимизације губитка. Она је дефинисана тако да рачуна разлику између предвиђених класа изведеног модела и предвиђених класа стварног модела. Решавањем ове линеарне функције за одређивање вероватноћа помоћу којих ће се променити предвиђене класе модела ради постизања изједначених шанси уз минимизацију губитка добија се праведан модел одлучивања [10].

3. МЕТОДОЛОГИЈА

Као циљ рада постављена је цена праведности модела машинског учења. Односно, потребно је измерити колика је промена тачности модела машинског учења након примене техника за постизање праведних модела.

Поставка експеримента. Пре примене техника биће изграђена два модела који ће служити за поређење. То су логистичка регресија () (енгл. *Logistic Regression*) и алгоритам случајних стабала одлучивања () (енгл. *Random Forest*). Ова два алгоритма су изабрана зато што се сматрају за најупотребљивије алгоритме у области откривања законитости у подацима и машинског учења [27].

Логистичка регресија је алгоритам класификације који се користи за предвиђање циљне варијабле која је категоричког типа. За разлику од линеарне регресије која даје

континуиране вредности бројева, логистичка регресија трансформише свој излаз користећи логистичку сигмоидну функцију да би вратила вредност вероватноће која се затим може пресликати у две или више дискретних класа. Сигмоидна функција је формулисана на следећи начин:

$$\sigma(t) = \frac{1}{1+exp^{(-t)}} \quad (9)$$

Одређивањем прага класификације, вероватноће се претварају у класу и добијају се резултати предикција [26].

Алгоритам случајних стабала одлучивања садржи у себи велики број појединачних стабала одлучивања која делују као целина. Свако појединачно стабло одлучивања предвиђа класу и класа са највише гласова постаје предвиђање модела [6]. Како је у раду [6] наведено битно је да модели стабла одлучивања буду међусобно некорелисани, а то се постиже тренирањем модела на различитим подскуповима атрибута. Модел се тиме штити од индивидуалних грешака појединачних стабала, где уколико једно погреши, остала ће својим предикцијама ту грешку исправити.

Модели ће бити валидирани дељењем скупа података на два подскупа, један за тренирање модела машинског учења и други за тестирање модела машинског учења. Тренинг подскуп чини 70%, док је тест 30% укупног броја опсервација у скупу података [23]. У овом раду се бирао праг који максимизира тачност, како би се одредила минимална цена повећања праведности модела. Та цена може бити и виша уколико се изабере праг којим модел постаје још праведнији.

Након утврђене полазне основе, биће примењене технике за постизање праведних модела машинског учења и то технике претпроцесирања уклањање различитог утицаја () и праведно **отежавање инстанци** (). [9] мења иницијални скуп података тако да није могуће разликовати дискриминисану и привилеговану групу у подацима за сваки посматрани атрибут. [12] не мења почетни скуп података, већ инстанцама додаје тежине. Модел у фази учења узима у обзир израчунате тежине чиме исправља дискриминаторно понашање модела. Наведене технике служе да уклоне дискриминација из тренинг скупа података [13]. На тај начин модел учи на скупу података који не указују на дискриминацију, што би за последицу требало да има праведан модел.

Поред техника претпроцесирања, приказаће се и примена технике постпроцесирања, које на већ наученом моделу коригују предикције [10]. Тачније примењене су класификација заснована на опцији одбацивања () и изједначавање прилика (). [12] уводи критичну зону око прага класификације у којој инстанце које припадају дискриминисаној и привилегованој групи буду класификоване у позитивну и негативну класу, тим редоследом. Тај исход се додељује без обзира да ли инстанаца из привилеговане групе има већу вероватноћу исхода од инстанце из дискриминисане групе, чиме се смањује дискриминација. Циљ технике OP [10] јесте постизање једнакости разлике у приликама између привилеговане и дискриминисане групе за већ научени бинарни класификатор, уз минимизацију губитка који се огледа у смањењу тачности модела.

Скупови података. Први скуп података који ће се обрађивати је Одрасли (енгл. *Adult*) скуп података [17]. Овај скуп података представља проблем бинарне класификације где је потребно на основу особина појединца, као што су демографске карактеристике, образовање и друге личне карактеристике предвидети да ли појединац зарађује на годишњем нивоу изнад или испод 50.000 долара (тзв. цензус). Одрасли представља типичан пример нежељене дискриминације, где мушкарци зарађују знатно више од жена. Скуп података се састоји из 45.222 опсервације и 44 улазна атрибута.

Други скуп података представља податке о пласирању студената кампуса приликом тражења посла, тзв. регрутовање студената (енгл. *Campus Recruitment*) [3]. Такође, и овај скуп података се бави проблемом бинарне класификације. Овај скуп података је знатно мањи него претходни и има укупно 215 опсервација. Такође, постоји дискриминација у погледу пола.

Мере тачности и праведности. За сваки скуп података и за сваку технику остваривања праведности користиће се мере подељене у две групе. Прва група служи да испита колико је модел машинског учења тачан, односно колико добро решава проблем који је постављен. За те потребе користе се балансирана тачност, ROC-AUC, прецизност и одзив.

Балансирана тачност класификације: Балансирана тачност (енгл. *Balanced accuracy*) представља просек стопе тачно предвиђених позитивних опсервација и стопе тачно предвиђених негативних опсервација [24].

$$BA = \frac{TPR+TNR}{2} \quad (10)$$

Прецизност: Прецизност (енгл. *Precision*) представља удео тачно предвиђених позитивних опсервација у опсервацијама које су предвиђене као позитивне [11].

$$Prec = \frac{TP}{TP+FP} \quad (11)$$

Одзив: Одзив (енгл. *Recall*) представља удео тачно предвиђених позитивних опсервација у опсервацијама које су стварно позитивне [11].

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

ROC-AUC: ROC-AUC (у даљем тексту) представља меру перформанси класификационог модела за различите вредности прага класификације. Вредности се крећу од 0 до 1 и указују на то колико је модел способан да направи разлику између класа које се предвиђају. ROC крива се добија исцртавањем стопе тачно предвиђене позитивне класе (енгл. TPR – *True Positive Rate*) на у оси и стопе погрешно предвиђене позитивне класе (енгл. FPR – *False Positive Rate*) на х оси, за различите вредности прага класификације [24].

Друга група мера служи да измери праведност одлука модела машинског учења. Коришћене су мере које су описане у прегледу литературе. То су различит утицај, разлика статистичких паритета, једнаке разлике у приликама и просечна разлика у приликама.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

Како су експерименти спроведени на два скупа података, користећи два алгоритма и две групе мера, резултати ће бити приказани посебно за сваки скуп података. Прво су приказани резултати за скуп података *Одрасли* у табели 1. Најбоље вредности по мерама су приказане подебљаним словима.

Табела 1. *Резултати примена техника праведности на скупу података Одрасли*

Алг.	Техника	Мере тачности				Мере праведности			
		BA	*AUC*	*Recall*	*Prec*	*DI*	*SP*	*EO*	AVG_0
LR	/	0,8263	0,9084	**0,8854**	0,5549	0,3003	-0,3560	-0,1208	-0,1879
	DIR	0,8128	0,8986	0,8313	0,5710	0,2896	-0,3318	-0,1571	-0,1893
	RW	0,8172	0,9000	0,8419	0,5709	0,5331	-0,2002	0,0500	-0,0235
	ROC	0,7925	0,9084	0,7738	0,5731	0,8253	**-0,0617**	0,1695	0,1050
	OP	0,7353	0,9084	0,5562	0,6805	0,6062	-0,0910	**0,0089**	**0,0062**
RF	/	**0,8298**	**0,9140**	0,8634	0,5812	0,2894	-0,3381	-0,1425	-0,1850
	DIR	0,8183	0,9010	0,8339	0,5818	0,2736	-0,3364	-0,1452	-0,1858
	RW	0,8231	0,9081	0,8612	0,5678	0,5127	-0,2165	0,0312	-0,0399
	ROC	0,7971	**0,9140**	0,8159	0,5468	**0,8358**	-0,0639	0,1386	0,0920
	OP	0,7044	**0,9140**	0,4711	**0,7127**	0,5421	-0,0877	-0,0154	-0,0095

Почетне мере евалуације показују да су оба алгоритма научили моделе који имају веома добре резултате мера тачности, са знатно израженом дискриминацијом. Односно, мере праведности нису у задовољавајућем опсегу. Другим речима, модели су тачни, међутим показују дискриминаторно понашање. Такође, можемо видети да је алгоритам насумичних стабала одлучивања био тачнији од логистичке регресије, са бољим вредностима балансиране тачности, AUC-а и прецизности. Детаљнија дискусија по техникама је приказана испод.

Уклањање различитог утицаја. Уклањање различитог утицаја није постигло значајне резултате у погледу постизања веће праведности модела. Приказани су резултати након примене алгоритма где је параметар ниво поправке једнак 1. Све метрике праведности су остале приближно једнаке почетним вредностима, чак је и приметан лошији резултат у односу на почетни модел. Из тог разлога у оваквој поставци техника уклањање различитог утицаја не би била повољан избор.

Отежавање инстанци. Ова техника није у потпуности одстранила различит утицај, али је смањила неправедност према женском полу. И даље мушкарци имају вишу стопу позитивних исхода у односу на жене, међутим у потпуности је уклоњена дискриминација у погледу једнакости прилика. Смањена је неправедност у погледу веће стопе погрешне класификације жена које имају високу плату у односу на мушкарце.

Минимална цена овог повећања праведности логистичке регресије износи 0,0091 уколико се посматра AUC, а уколико се узме у обзир балансирана тачност 0,0084. То је веома мала цена за резултате који су постигнути, те то оставља простора за додатно спуштање тачности у циљу повећања праведности модела. Алгоритам насумичних стабала одлучивања је дао лошије резултате у погледу праведности, па је и минимална цена која износи 0,0067 за AUC нижа, у односу на логистичку регресију. Изједначене су разлике у приликама, као и у претходном случају док су вредности метрика различит утицај и разлика статистичких паритета лошији.

Класификација заснована на опцији одбацивања. Метрике праведности су након примене ове технике дале задовољавајуће резултате. Различит утицај је 0,8253 што представља праведно поступање према припадницима оба пола. Такође разлика статистичких паритета је у опсегу дозвољених вредности. Интересантно је приметити да је, при минимизацији цене повећања праведности, дискриминација према женама смањила у тој мери да је модел почео да дискриминише мушки пол. Тако су вредности метрика које указују на једнакост прилика група неповољне за мушкарце. Ова појава се може посматрати као позитивна дискриминација [20].

Минимална цена за ове резултате праведности износи смањење балансиране тачности у износу од 0,0338 за модел логистичке регресије. За око 3% би модел више грешио да би избегао дискриминацију. Тачност модела је, у случају модела насумичних стабала одлучивања, смањена за 0,0327, што на стопу тачности од 83% не представља велики проблем, док су резултати постигнути на пољу праведности слични.

Изједначене прилике. Ова техника је у погледу цене најскупља од свих које су примењене у овом раду. За постизање веће праведности модела балансирана тачност се смањила за 0,0910. Тачност од 73,53% након примене технике може бити прихватљива и то у највећој мери зависи од проблема који се решава, међутим, у случају када се лако постиже виша тачност, овај резултат није задовољавајући. Приметан је и пад одзива. Наиме, након примене технике повећава се број опсервација којима је модел погрешно доделио негативну класу за 30%. То су погрешно класификовани мушкарци, како би се уједначио однос мушкараца и жена са високим примањима и тиме модел постао мање дискриминишући.

За крајњу одлуку да ли је примена ове технике оправдана, мора се узети у обзир проблем који се обрађује и колико је битно постићи праведан модел.

Резултати за примену ове технике на алгоритму насумичних стабала одлучивања су лошији него у случају алгоритма логистичке регресије. Минимална цена примене ове технике износи 12,54% смањења балансиране тачности, те је закључак да је ову технику, уколико се она одабере у складу са проблемом, боље примењивати са једноставнијим класификатором, тј. логистичком регресијом.

Табела 2. *Резултати примена техника праведности на скупу података Регрутовање студената*

Алг.	Техника	Мере тачности				Мере праведности			
		BA	*AUC*	*Recall*	*Prec*	*DI*	*SP*	*EO*	AVG_0
LR	/	0,8172	0,8898	**0,9677**	0,8824	0,8928	-0,0881	-0,0909	-0,1080
	DIR	**0,9286**	0,9036	0,8571	**1,0000**	0,7862	-0,1586	-0,1068	-0,0534
	RW	0,8172	0,8898	**0,9677**	0,8824	0,8928	-0,0881	-0,0909	-0,1080
	ROC	0,8105	0,8898	0,8710	0,9000	1,0807	0,0548	0,0591	0,0295
	OP	0,8172	0,8898	**0,9677**	0,8824	0,8928	-0,0881	-0,0909	-0,1080
RF	/	0,7433	0,8656	0,9032	0,8485	0,9333	-0,0524	0,0091	-0,1205
	DIR	0,8661	**0,9214**	0,8571	0,9677	1,0568	0,0403	0,0427	0,1880
	RW	0,7016	0,8575	0,9032	0,8235	**1,0182**	**0,0143**	**0,0091**	**0,0045**
	ROC	0,7876	0,8656	0,7419	0,9200	0,1920	1,4667	0,2333	0,2591
	OP	0,6344	0,8656	0,7838	0,9355	0,8960	-0,0929	-0,0409	-0,1455

Специфичност скупа података Регрутовање студената у кампусу јесте веома мали број опсервација. Када скуп података садржи мало опсервација над којима модел може да учи, врло лако може доћи до модела машинског учења који је пренаучен. Међутим, и поред малог броја опсервација, логистичка регресија се показала јако добро на овом скупу података.

За разлику од претходног скупа података где је алгоритам насумичних стабала одлучивања био бољи од алгоритма логистичке регресије, овде то није случај. Алгоритам насумичних стабала одлучивања је показала знатно лошије резултате у погледу тачности, али нешто боље у погледу праведности модела.

Иако је модел без примењених техника добар, над њим се могу применити технике како би се покушао смањити обим дискриминације који тренутно постоји. На тај начин могу се приближити метрике праведности њиховим идеалним вредностима, што је и био циљ овог експеримента.

Уклањање различитог утицаја. Примена ове технике се за модел логистичка регресија није добро показала. Приметан је пад вредности мере различит утицај тако да је постигнут супротан ефекат, модел је научио да дискриминише са измењеним скупом података. Међутим приметан је повољан утицај у случају модела насумичих стабала одлучивања. Мере праведности су ближе идеалним вредностима, док су мере тачности и веће у односу на почетне. Могуће је да је дошло до оваквог резултата услед ограниченог броја опсервација.

Отежавање инстанци. У случају логистичке регресије ова техника није имала никаквог утицаја на промену вредности метрика. Метрике тачности и метрике праведности су, за праг класификације који је изабран, остале једнаке.

У случају насумичних стабала одлучивања, ова техника је имала ефекта. Све метрике праведности су сведене на готово идеалну вредност, односно скоро да уопште не постоји дискриминација. Минимална цена примене ове технике износи 0,0081 за AUC и 0,0417 у случају балансиране тачности. У овом случају треба преиспитати да ли је потребно имати приближно идеалне метрике праведности.

Класификација заснована на опцији одбацивања. У случају логистичке регресије постигнут је скоро па идеалан резултат. Минимална цена за ове резултате праведности износи смањење балансиране тачности у износу од 0,0067. За мање од 1% би модел више грешио да би елиминисао дискриминацију, што се сматра веома добрим балансом између тачности и праведности модела машинског учења.

За насумична стабла одлучивања ова техника је постигла веома лоше резултате. Тачност модела се повећала, међутим све метрике праведности показују високу дискриминацију према мушкарцима. Посматрајући критичну зону, може се уочити да је она била превелика у односу на димензионалност скупа података, те се додељивала негативне класе свим мушкарцима како би се смањила дискриминација према женама. То је довело до тога да модел дискримише мушкараце.

Изједначене прилике. У овом експерименту се ова техника примењена над насумичним стаблима одлучивања најлошије показала, док су вредности метрика за логистичку регресију остале непромењене.

5. ЗАКЉУЧАК

Овај рад се бавио проблемом дискриминације у примени алгоритама машинског учења. Циљ рада, поред примене техника за смањење дискриминације у моделима, је био и измерити колика је цена њихове примене. Односно колико мора да се жртвује тачност модела машинског учења како би се постигла праведност. Цена је дефинисана као проценат смањења метрика тачности за повећање метрика праведности модела.

Изведена су два експеримента. Први на скупу података Одрасли и други на скупу података Регрутација студената у кампусу. На оба скупа података испробана су два алгоритма која су међу најпопуларнијим у области машинског учења. То су логистичка регресија и алгоритам насумичних стабала одлучивања. Прво су измерене почетне вредности мера тачности и праведности које су служиле као полазна основа за поређење, тј. балансирана тачност, AUC, одзив и прецизност као мере тачности и различит утицај, разлика статистичких паритета, једнаке разлике у приликама и просечна разлика у приликама као мере праведности. Након тога, примењене су две технике претпроцесирања – уклањање различитог утицаја и отежавање инстанци, као и две технике постпроцесирања класификација заснована на опцији одбацивања и изједначене прилике. Након примене сваке од техника модели су евалуирани и мерена је њихова тачност и праведност.

Цена примене техника за постизање праведних модела варира у зависности од скупа података и модела машинског учења који се тренира. Технике претпроцесирања и постпроцесирања могу да постигну боље резултате метрика праведности, међутим то не мора увек бити случај. Други експеримент је показао исте и лошије метрике праведности у односу на почетне резултате модела. У оба експеримента се једноставнији алгоритам, тј. логистичка регресија показала као бољи модел у погледу оба циља. Експерименти су показали да су технике претпроцесирања рачунски мање сложене у односу на технике постпроцесирања, међутим технике постпроцесирања постижу боље резултате у погледу праведности. Од техника претпроцесирања отежавање

инстанци се показало као боља опција у оба случаја, док је класификација заснована на опцији одбацивања постигла боље резултате у погледу постпроцесирања. У зависности од проблема који се решава, приоритет се може дати циљу тачности или циљу праведности и у складу са тим вршити одабир технике која ће се применити.

У даљим истраживањима обратиће се пажња на прилагођавање алгоритама машинског учења како би се постигли праведни модели. Овај приступ може да да гаранције постизања праведних модела кроз додавање ограничења у математички модел [21]. Такође, прилагођавањем алгоритма омогућава се уградња производне мере праведности, те се алгоритам може прилагодити проблему који се решава.

ЛИТЕРАТУРА

[1] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*.

[2] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

[3] Ben Roshan (2020). Kaggle. Campus Recruitment [https://www.kaggle.com/benroshan/factorsaffecting-campus-placement]

[4] Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency* (pp. 149-159). PMLR.

[5] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[6] Breiman, L. (1999). Random forests. UC Berkeley TR567.

[7] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806).

[8] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214-226).

[9] 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).

[10] Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, *29*, 3315-3323.

[11] Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1.

[12] Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1-33.

[13] 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.

[14] Kamishima, T., Akaho, S., & Sakuma, J. (2011, December). Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 643-650). IEEE.

[15] Karim, S., & Beardsley, K. (2017). *Equal opportunity peacekeeping: women, peace, and security in post-conflict states*. Oxford University Press.

[16] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019, January). An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 100-109).

[17] Kohavi, R. (1996, August). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *International Conference on Knowledge Discovery & Data Mining* (Vol. 96, pp. 202-207).

[18] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

[19] Murphy, A., Kelly, B., Bergmann, K., Khaletskyy, K., O'Connor, R. V., & Clarke, P. M. (2019, September). Examining Unequal Gender Distribution in Software Engineering. In *European Conference on Software Process Improvement* (pp. 659-671). Springer, Cham.

[20] Noon, M. (2010). The shackled runner: time to rethink positive discrimination?. *Work, Employment and Society*, *24*(4), 728-739.

[21] Radovanović, S., Petrović, A., Delibašić, B., & Suknović, M. (2020, August). Enforcing fairness in logistic regression algorithm. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-7). IEEE.

[22] Sharifi-Malvajerdi, S., Kearns, M., & Roth, A. (2019). Average Individual Fairness: Algorithms, Generalization and Experiments. In *Advances in Neural Information Processing Systems* (pp. 8242-8251).

[23] Suknović, M., & Delibašić, B. (2010). Poslovna inteligencija i sistemi za podršku odlučivanju. *FON, Beograd*.

[24] Tharwat, A. (2020). Classification assessment methods. Applied Computing and Informatics.

[25] Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 2053951717743530.

[26] Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (p. 217–244). American Psychological Association.

[27] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.

[28] Yapo, A., & Weiss, J. (2018, January). Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

[29] Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. *J. Mach. Learn. Res.*, *20*(75), 1-42.

[30] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335-340).

Милица Перишић, Универзитет у Београду – Факултет организационих наука
Контакт: milicaperisic507@gmail.com
Област интересовања: Big Data, машинско учење, анализа података

Сандро Радовановић, асистент, Универзитет у Београду – Факултет организационих наука
Контакт: sandro.radovanovic@fon.bg.ac.rs
Област интересовања: Машинско учење, системи за подршку одлучивању, теорија одлучивања, откривање законитости у подацима

др Борис Делибашић, редовни професор, Универзитет у Београду – Факултет организационих наука
Контакт: boris.delibasic@fon.bg.ac.rs
Област интересовања: Системи за подршку одлучивања, развој алгоритма машинског уђења, пословна интелигенција, теорија одлучивања