VEŠTAČKA INTELIGENCIJA I MAŠINSKO UČENJE – TRENDOVI RAZVOJA ETIČKIH PRIMENA

Nikola Gradojević[1], Nebojša Ralević[2], Vladimir Đaković[3]
[1,2,3]*Univerzitet u Novom Sadu, Fakultet tehničkih nauka, Novi Sad, Srbija*
[1]ngradoje@uns.ac.rs, [2]nralevic@uns.ac.rs, [3]v_djakovic@uns.ac.rs

Kratak sadržaj: *Predmet istraživanja u ovom radu je analiziranje izazova i mogućih problema koje primena veštačke inteligencije i mašinskog učenja neminovno donose. Cilj istraživanja predstavlja identifikovanje postojanja diskriminacije u primeni veštačke inteligencije i mašinskog učenja u obrazovanju sa mogućnostima prevazilaženja uočenih nedostataka, i to naročito kod prediktivnih i klasifikacionih modela. Rezultati istraživanja u radu su od značaja kako profesionalnoj javnosti, tako i kreatorima politika u predmetnoj oblasti, naročito u funkciji dijagnostikovanja potencijalne diskriminacije i neetičnosti u primeni veštačke inteligencije i mašinskog učenja.*

Ključne reči: Veštačka inteligencija, AI, mašinsko učenje, ML, diskriminacija, obrazovanje, etika.

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING – TRENDS IN DEVELOPING ETHICAL APPLICATIONS

Abstract: *The research objective of this paper is to analyze the challenges and potential problems that artificial intelligence and machine learning applications may inevitably involve. This research aims at identifying the existence of discrimination in the applications of artificial intelligence and machine learning in education with the possibility of overcoming the possible shortcomings, especially in the case of predictive and classification models. The results of this research will be valuable both to the professional community and to government policy-makers in the subject area, especially in detecting potential ethical biases and discrimination in the applications of artificial intelligence and machine learning.*

Key Words: Artificial Intelligence, AI, Machine Learning, ML, Discrimination, Education, Ethics.

1. UVOD

Dinamizam promena u okruženju uz imperativ implementacije alata digitalnog doba u velikoj meri determinišu oblikovanje inženjerskih studijskih programa, i to naročito u oblasti tehničko-tehnoloških nauka. Fakultet tehničkih nauka Univerziteta u Novom Sadu (FTN) jeste lider u teorijsko-aplikativnoj primeni savremenih inženjerskih metoda, tehnika i alata, koji korespondiraju potrebnim kompetencijama studenata u svakodnevnoj poslovnoj praksi [1].

U tom smislu, posebno se izdvaja inkorporiranje veštačke inteligencije (engl. *Artificial Intelligence* – AI) i mašinskog učenja (engl. *Machine Learning* – ML) u studijske programe orijentisane na rešavanje kompleksnih stručno-aplikativnih inženjerskih izazova i donošenje optimalnih poslovnih odluka, a u funkciji daljeg rasta i razvoja i sticanju i održanju konkurentske prednosti preduzeća. Poseban kvalitet predstavlja multidisciplinarnost Fakulteta tehničkih nauka, koja se ogleda i u saradnji većeg broja Katedri i Departmana, uz učešće istraživača orijentisanih na primenu AI i ML-a.

Studentima Fakulteta tehničkih nauka se omogućava savladavanje teorijskih i aplikativnih osnova AI i ML-a kako u okviru nastavnog procesa, tako i mogućnosti učestvovanja na praktično orijentisanim naučno-istraživačkim projektima, koji omogućuju izvrsnost u predmetnoj oblasti. Data komparativna prednost Fakulteta tehničkih nauka se ogleda i u mogućnosti studenata da uz mentorski rad vrše dodatno usmeravanje shodno preferencijama na osnovnim akademskim studijama, a naročito na master akademskim i doktorskim akademskim studijama. Horizontalna i vertikalna povezanost čini optimalni miks koji u značajnoj meri omogućava studentima da steknu odgovarajuće kompetencije i znanja u predmetnoj oblasti, koja će odražavati distinktivnu kompetentnost u skladu sa posebnim preferencijama svakog studenta ponaosob. To se omogućava i posedovanjem odgovarajuće naučno-istraživačke baze na Fakultetu tehničkih nauka na svim nivoima studija.

Rešavanje problema u poslovnoj praksi, naročito indukovanih frekventnom pojavom kriznih stanja i ekstremnih događaja, zahtevaju generisanje odgovarajućih algoritama koji će sa protokom vremena moći da poboljšavaju performanse sa ciljem maksimizacije efekata i prevazilaženjem ograničenja. Poseban izazov predstavlja primena AI i ML-a kod predikcionih modela. Dati modeli koriste velike setove podataka koji sadrže osnovu za kreiranje odgovarajućeg modela, odnosno razvijanje predikcionog algoritma. Poznato je da se predikcioni modeli baziraju na setovima istorijskih podataka, koji mogu biti pristrasni i nepotpuni ili čak sadržati prošle diskriminatorne odluke. Sledstveno, uspešnost istih korelira sa mogućnošću konstantnog inoviranja korišćenih

podataka, jer je učenje kontinualni proces.

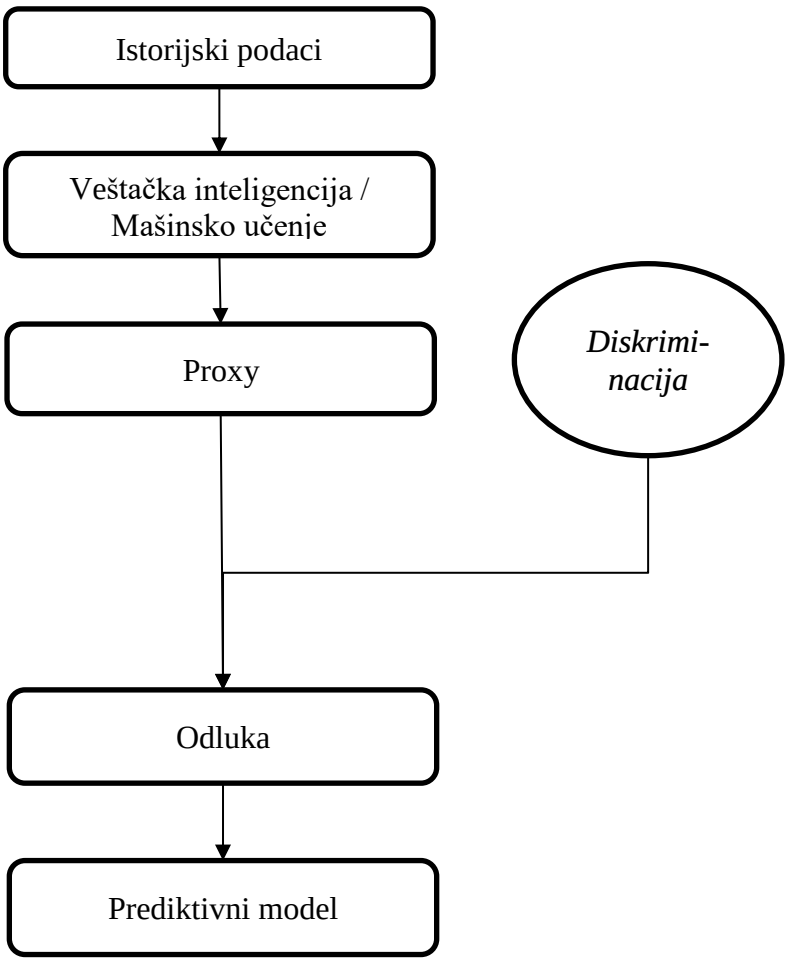
Prvi cilj ovog rada je da ukaže na mogućnost da upotreba AI i ML-a u obrazovanju donosi i određene rizike i probleme. Na primer, kompanija Amazon je utvrdila da je njihov klasifikacioni algoritam u odlukama zapošljavanja konstantno favorizovao kandidate muškog pola, a diskriminisao je kandidate ženskog pola. U kontekstu obrazovanja, neki od sličnih izazova uključuju neetičko bodovanje kandidata na prijemnom ispitu za fakultet gde algoritam može da vrši sistematsku diskriminaciju na bazi nacionalnosti, rase, pola, vere ili čak mesta rođenja prijavljenih kandidata [2].

Drugi važan cilj ovog istraživanja je da predloži moguća rešenja problema diskriminacije i neetičke primene AI. Jedno od rešenja je uvođenje adekvatnih zakona koji bi propisali da svaka organizacija (što uključuje i institucije obrazovanja) mora da sprovodi redovnu internu reviziju svojih algoritama AI koji donose automatske odluke u cilju otkrivanja i otklanjanja diskriminacionih odluka. Dalje, rešenje problema sugeriše i da ljudski faktor mora biti zastupljeniji u donošenju odluka kao komplement mašinskom odlučivanju. Konačno, problemu je moguće pristupiti i sa tehničkog stanovišta putem unapređenja algoritama učenja AI, unapređenja kvaliteta podataka, metoda učenja (savremenim neuronskim mrežama; npr. ***transformer*** mreže), kao i hibridizacijom sa srodnim metodama kao što je ***fuzzy*** logika i modelima na bazi „stabala" odlučivanja (npr. ***random forest*** model).

2. DISKRIMINACIJA KAO IZAZOV U PRIMENI AI I ML-A

Uz sve prednosti koje donosi primena AI i ML-a, nameću se pitanja etike i morala, a naročito u pogledu diskriminatornih mera sa kojima se može susresti prilikom primene istih [3]. Najčešći slučaj je prilikom korišćenja velikih setova podataka koji se koriste prilikom primene AI i ML-a (tzv. ***big data***) [4]. Ukoliko je u pitanju pristup crne kutije, donošenje odluka se odvija na način da nije moguć uvid u način korišćenja podataka, formiranje i učenje algoritama, a konsekventno i uspešnost primene, što implicira da je za takav način neophodno ustanovljavanje zaštićenih varijabli ili osetljivih atributa, u zavisnosti od oblasti primene (finansije, osiguranje, zapošljavanje, procena rizika i sl.) [5], [6].

Na slici 1. može se videti uzročno-posledični sled u primeni AI i ML-a u funkciji kreiranja prediktivnog modela sa posebnim akcentom na diskriminativne varijable.



Slika 1. Diskriminacija u primeni AI i ML-a [Autori]

Zaštićene karakteristike određene varijable mogu biti bilo binarne, kategorične ili numeričke, u zavisnosti od same prirode korišćenih podataka, koje je potrebno zaštititi, kako ne bi došlo do diskriminacije. Ukoliko je slučaj o mogućnosti diskriminacije po više osnova, onda je reč o višestrukoj diskriminaciji. Takođe, u zavisnosti od prirode zaštićenih karakteristika koje se koriste kao ulaz (input) u AI model, može se govoriti o direktnoj ili indirektnoj diskriminaciji.

Types of proxy discrimination	Definition	Is Suspect variable Directly or Indirectly predictive?	Risk of Proxy Discrimination by AI
Causal proxy discrimination	Legally-suspect characteristic (i.e. race, genetics, health) causally linked to target variable (i.e. expected insurance costs).	Directly Predictive, as suspect variable contains predictive power that cannot be more directly captured by facially-neutral data.	Very high risk as AIs will inevitably proxy for suspect characteristic.
Opaque proxy discrimination	Legally-suspect characteristic (i.e. race, genetics, health) predictive of target variable (i.e. expected insurance costs) for reasons not mediated through a presently quantifiable or available variable.	Directly Predictive, as suspect variable contains predictive power that cannot be more directly captured by facially-neutral data.	High risk as AIs will inevitably proxy for suspect characteristic until better data or causal mechanism becomes available.
Indirect proxy discrimination	Legally-suspect characteristic (i.e. race, genetics, health) predictive of target variable (i.e. expected insurance costs) because it proxies for a quantifiable or available variable.	Indirectly Predictive, as suspect variable only contains predictive power because it proxies for another, quantifiable and potentially available, variable that is not included in the AI's training data.	Moderate risk as AIs will only proxy discriminate if (i) data on causative facially-neutral characteristic is not available, and (ii) better proxies for causative characteristic than suspect characteristic are not available.

Slika 2. Tipovi proxy diskriminacije [7], prilagođeno.

Tipovi proxy diskriminacije sa mogućnošću direktne ili indirektne predikcije su dati na slici 2.

3. NASTAVA IZ AI I ML-A NA FTN-U

Na velikoj većini univerziteta u svetu se izučava AI, dok na Fakultetu tehničkih nauka Univerziteta u Novom Sadu postoji više studijskih programa gde se održava nastava iz predmeta koji su direktno ili indirektno povezani sa njom. Veštačka inteligencija i mašinsko učenje je studijski program master akademskih studija kompletno posvećen toj temi (FTN, akreditacija 2020.). Jedan od autora predaje predmete iz date oblasti na studijskom programu Matematika u tehnici dvogodišnjih master akademskih studija (FTN, akreditacija 2020.).

Predmeti direktno povezani sa veštačkom inteligencijom na studijskom programu Matematika u tehnici su:

1) Računarska inteligencija – matematičke osnove,
2) Mašinsko učenje u ugrađenim sistemima,
3) Osnovi mašinske vizije,
4) Mašinsko učenje,
5) Matematički modeli u računarskoj viziji,
6) Matematičke osnove prepoznavanja oblika.

Nastava iz AI se održava i delimično na osnovnim akademskim studijama. Na doktorskim akademskim studijama sem predmeta gde se izlažu najnovija dostignuća daje se i originalan doprinos kroz veliki broj naučnih radova, ali i kroz rad na domaćim i međunarodnim projektima.

U skladu sa trenutnim trendovima razvoja AI, u budućnosti bi trebalo uvesti dodatne predmete na osnovnim i postdiplomskim studijama koji uključuju tematiku etike i diskriminacije u AI, kao i upravljanje politikom AI, praćenjem usklađenosti sa zakonom, upravljanje rizikom, procesima donošenja odluka, izborom metrike i izveštavanjem, kao i revizijom i praćenjem AI.

4. ZAKLJUČAK

Sistemi AI u obrazovanju (a i šire) koji obezbeđuju performanse sistema AI i njihovu etičku upotrebu moraju uspostaviti kontrolne parametre kao što su tačnost, preciznost, opoziv (*recall*), stepen ljudskog odlučivanja i distribucija zaštićenih grupa po verskoj, polnoj, rasnoj ili sličnim potencijalno diskriminativnim osnovama.

Ovakva politika je od suštinskog značaja za obezbeđivanje da se AI tehnologija koristi u okviru etičkih, zakonskih i granica željenih performansi. Proces evaluacije AI politike uključuje utvrđivanje statusa rizika i statusa usklađenosti AI sistema na osnovu različitih ulaznih podataka kao što su rezultati testova validnosti podataka i metapodataka, ljudski inputi/odobrenja, operativni parametri i kontekstualne informacije u skladu sa predviđenim performansama sistema.

Korišćenjem aktivnih *in-house* AI sistema upravljanja, organizacija može da pojednostavi usklađenost i postigne minimalne garancije učinka dok sa lakoćom ispunjava regulatorne i interne obaveze. Nužno je da ovakav trend razvoja primene AI i ML-a prati obrazovanje na FTN-u, ali i na ostalim obrazovnim institucijama u Republici Srbiji.

ZAHVALNICA:

Rad je podržan od strane projekta „Unapređenje nastavnog procesa na engleskom jeziku u opštim disciplinama", Departmana za opšte discipline u tehnici, Fakulteta tehničkih nauka, Univerziteta u Novom Sadu, kao i od strane Ministarstva nauke, tehnološkog razvoja i inovacija kroz projekat broj 451-03-47/2023-01/200156 „Inovativna naučna i umetnička istraživanja iz domena delatnosti FTN-a".

5. LITERATURA

[1] Nebojša M. Ralević, Vladimir Đ. Đaković, *Matematika u tehnici: primena inovativnih inženjerskih metoda, tehnika i alata*. XXIX Skup Trendovi Razvoja: "Univerzitet pred novim izazovima", Vrnjačka Banja, 08.-11.02.2023. Univerzitet u Novom Sadu, Fakultet tehničkih nauka, str. 56-59, 2023.

[2] Joseph L. Breeden, Eugenia Leonova, *Creating Unbiased Machine Learning Models by Design*, Journal of Risk and Financial Management, 14 (11): 565, pp.1-15, 2021, https://doi.org/10.3390/jrfm14110565.

[3] Bert Heinrichs, *Discrimination in the age of artificial intelligence*, *AI & Society*, pp. 1-12, 2022.

[4] Talia B. Gillis, Jann L. Spiess, *Big data and discrimination*, The University of Chicago Law Review, 86(2), pp. 459-488, 2019.

[5] Sumit Das, Aritra Dey, Akash Pal, Nabamita Roy, *Applications of artificial intelligence in machine learning: review and prospect*, International Journal of Computer Applications, 115(9), pp. 31-41, 2015.

[6] Philippe Bracke, Anupam Datta, Carsten Jung, Shayak Sen, *Machine learning explainability in finance: an application to default risk analysis*, Bank of England, 816, pp. 1-43, 2019.

[7] Anya E.R. Prince, Daniel Schwarcz, *Proxy discrimination in the age of artificial intelligence and big data*, Iowa Law Review, 105, pp. 1257-1318, 2019.