Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva, Zavod za komunikacijske i svemirske tehnologije
Filozofski fakultet, Odsjek za psihologiju

Barbara Geld
Rea Pintar

Klasifikacija studenata prema znanstvenom području kojem pripadaju na temelju iskazanih stavova i navika metodama strojnog učenja

Zagreb, lipanj 2022.

Ovaj rad izrađen je na Zavodu za komunikacijske i svemirske tehnologije Fakulteta elektrotehnike i računarstva te Odsjeku za Psihologiju Filozofskog fakulteta Sveučilišta u Zagrebu, pod vodstvom doc. dr. dc. Daria Bojanjca i prof. dr. sc. Vesne Buško i predan je na natječaj za dodjelu Rektorove nagrade u
akademskoj godini 2021./22.

Sadržaj

Uvod

Prema Akertu, Aronsonu i Wilsonu (2005; prema Eagly i Chaiken, 1993, 1998; Olson i Zanna, 1993; Petty i Wegener, 1998; Petty, Wegener i Fabrigar, 1997) stav je definiran kao *vrednovanje ljudi, objekata ili ideja,* a konstrukt je kojime se primarno bavi socijalna psihologija. Već desecima godina znanstvenici, za porijeklo stavova, pružaju raznolike teorije. Prema Tesseru (1993; prema Akert i sur., 2005) te njegovim istraživanjima na jednojajčanim i dvojajčanim blizancima, nastanak stavova vezan je uz gene, budući da jednojajčani blizanci dovoljno konzistentno iskazuju sličnije stavove nego dvojajčani. Unatoč tim rezultatima, ni sam Tesser ne tvrdi da postoje specifični geni zaduženi isključivo za određivanje naših stavova. U svom istraživanju o važnosti heritabilnosti u psihologijskim istraživanjima, specifično u području stavova, Tesser (1993) spominje formulu o totalnoj fenotipskoj varijanci. Objašnjava kako se fenotip – izraženost određenih ponašanja, osobina pa i stavova – temelji na interakciji gena i okoline, kombinaciji faktora koja je u praksi za većinu ljudi potpuno različita. Posljedično, Tesser (1993) odbacuje tezu o specifičnim genima zaduženima za određivanje stavova te *stavove* predstavlja kao konstrukt koji jest, u većoj ili manjoj mjeri, pod utjecajem heritabilih efekata, ali istovremeno i utjecajem okoline u kojoj osoba raste i uči.

Utjecaj socijalnog iskustva koji prihvaća i Tesser, temeljni je pristup usvajanja i razvijanja stavova kojeg istražuju i kojime se bave socijalni psiholozi danas (Akert i sur., 2005). Proces formiranja stavova, prema Akertu i sur. (2005), može biti kognitivno, emocionalno ili ponašajno utemeljen, a ti se onda stavovi razlikuju i po svome porijeklu, ali i funkciji. Dok kognitivno zasnovani stavovi proizlaze iz racionalne, detaljne analize relevantnih karakteristika objekta stava, emocionalno zasnovane stavove vežemo uz osjećaje te subjektivni dojam koji pojedinac ima prema objektu stava. Iako je za neke stavove intuitivno, ili čak korisno pretpostaviti da su kognitivno zasnovani, kao što je odabir političkog kandidata kojemu ćemo dodijeliti glas u nadolazećim izborima, ipak se pokazuje da to nije slučaj. Wattenberg (1987; prema Akertu i sur., 2005) dokazuje kako stav oko odabira političkih kandidata, zbog očiglednog manjka informiranosti, a istovremenih pozitivnih osjećaja usmjerenih prema kandidatima, nije kognitivno, već emocionalno utemeljen stav. Izvorište emocionalno utemeljenih stavova nije u razmatranju činjenica, već u moralnim uvjerenjima koje je pojedinac usvojio kroz odrastanje pa sad primjenjuje na razna društveno-osjetljiva i ostala pitanja. Uzimajući u obzir kako porijeklo tako i funkciju emocionalno zasnovanih stavova, nedvojbeno je lako odrediti da su se u ovom istraživanju najvećim dijelom ispitivali stavovi koji su

emocionalno zasnovani. To potvrđuje i funkcija emocionalno zastupljenih stavova koju definiraju razni autori (Breckler i Wiggins, 1989; Zanna i Rempel, 1988; prema Akert i sur., 2005), a koja opisuje takve stavove kao subjektivno, osobno-korisno potvrđivanje onog sklopa vrijednosti koji već posjedujemo, radije nego novo stvaranje objektivne, točne slike svijeta.

Iako je u ovom istraživanju cilj i svrha bio ispitati stavove i navike studenata, inspiracija za istraživanje, a kasnije i konstrukciju originalnog upitnika, proizašla je iz istraživanja interesa na osnovnoškolskoj, srednjoškolskoj i studentskoj populaciji (Šverko, 2008) te tamo korištenog instrumentarija i modela. *Personal Globe Inventory* ili *PGI,* instrument je korišten pri mjerenju interesa (Tracey, 2021), a njegovi rezultati, kod Šverko (2008), u interpretaciji uz sebe vežu i sferni model (Prediger, 1982). Predigerov je model sastavljen primarno od triju aspekata posla za koje se interes mjeri, a to su *ljudi-stvari, podaci-ideje* te *prestiž*. Zbog podijeljenosti sfernog modela na kvadrante te pridruženih interesa koji iste opisuju, uvidjeli smo paralelu između onoga što opisuje model i konstrukata od interesa u našem istraživanju. Uočena je poveznica kvadranata s znanstvenim područjima te interesima i osobinama s aspektima svakodnevnog života. Vođene modelom, u ovom istraživanju, kao dimenzije koje se ispituju, odabrale smo teme bliske studentima današnjice, a kao varijablu po kojoj smo predviđali da će se ispitanici moći grupirati jesu znanstvena područja kojima studenti pripadaju. Dimenzije, u istraživanju postavljene kao subskale upitnika koje istražujemo jesu: *Glazba*, *Sport*, *Umjetnost i kultura*, *Ekologija*, *Tehnologija*, *Društvena pitanja i odgovornost* i *Vjerovanja*. Iako psihologija najčešće pribjegava konfirmatornim studijama, ovo istraživanje označava suradnju psihologijske i računarske znanosti te smo odlučile za eksploratorni pristup radu koji se očituje u metodi, pristupu i obradi podataka. Jedna od korištenih metoda u analizi podataka jest nadzirano strojno učenje. Općenito, strojno učenje ima sve veću primjenu u analizi podataka jer pruža dublji uvid u podatke te se radi o već implementiranim algoritmima u raznim programskim jezicima koji omogućuju bržu i efikasniju analizu, osobito na velikim skupovima podataka. Osim samih algoritama, u programskim jezicima poput *Python*-a postoje i knjižice (skupovi funkcija i algoritama) posebno namijenjene za vizualizaciju podataka na posebne načine (Subasi, 2020), kakve do sada nismo mogli vidjeti u jezicima i okružjima za statističko modeliranje poput jezika *R*.

S ciljem ispitivanja razlika među skupinama studenata s obzirom na znanstveno područje kojemu pripada njihov smjer studija, u ovom su se istraživanju ispitivali stavovi, ali i usko vezane navike studenata Sveučilišta u Zagrebu. Kao što je spomenuto, za većinu stavova o kojima se sudionike ispitivalo pretpostavlja se da su emocionalno uspostavljeni, a navike o kojima su se prikupljali podaci su karakteristična ponašanja vezana uz te stavove. Detaljnije,

cilj ovog istraživanja bio je utvrditi postoje li, i koje su to, sličnosti u stavovima i navikama skupine ljude koje veže studij u istom znanstvenom području ili na istom fakultetu. Budući da se zadovoljstvo na radnom mjestu veže uz osobne interese, vrijednosti i kompetencije, imali smo za cilj ispitati postoje li dijeljene navike te značajno slični stavovi o nekim općim, svakodnevnim temama kao i društveno osjetljivim pitanjima.

Također želimo ispitati korisnost i uspješnost modela strojnog učenja nad prikupljenim skupom podataka, što je ekvivalentno ispitivanju razlika među skupinama studenata s obzirom na područje studiranja vezanih uz njihove navike i stavove. Neki od primjera ispitivanja uspješnosti modela koje ćemo uspoređivati u ovom radu (Naivni Bayesov klasifikator, Stabla odluke, Logistička regresija i dr.) jesu istraživanja nad eksploratornim skupovima podataka u biomedicinskom području poput Framingham Heart Study (Mahmood, Levy, Vasan, i Wang, 2014), te Wisconsin Breast Cancer (Obaid, Mohammed, Ghani, Mostafa i Taha, 2018).

Pitanja i problemi

1. Razlikuju li se studenti i studentice Sveučilišta u Zagrebu prema navikama i stavovima koje iskazuju u određenim česticama dimenzija, s obzirom na to kojem znanstvenom području pripadaju?
2. Jesu li metode nadziranog strojnog učenja korisne u svrhu razlikovanja studenata s obzirom na znanstveno područje kojemu pripadaju, pri obradi eksploratornog skupa podataka temeljenog na rezultatima *Upitnika o navikama i stavovima studenata?*

1. Ispitati postoji li razlika između studenata Sveučilišta u Zagrebu s obzirom na to kojem znanstvenom području pripadaju, pri iskazivanju navika i stavova konstruiranim *Upitnikom o navikama i stavovima studenata*.

2. Ispitati korisnost metoda nadziranog strojnog učenja u svrhu razlikovanja studenata, na temelju znanstvenog područja kojemu pripadaju, pri obradi eksploratornog skupa podataka temeljenog na rezultatima *Upitnika o navikama i stavovima studenata.*

Hipoteze

1. Hipoteza: Postoji razlika između studenata Sveučilišta u Zagrebu s obzirom na to kojem znanstvenom području pripadaju, pri iskazivanju navika i stavova mjerenih *Upitnikom o navikama i stavovima studenata*.

2. Hipoteza: Korištene metode nadziranog strojnog učenja bit će korisne pri obradi rezultata dobivenih u *Upitniku o navikama i stavovima studenata*, u svrhu razlikovanja studenata na temelju znanstvenog područja kojemu pripadaju.
3. Hipoteza: Korištene metode nadziranog strojnog učenja razlikovat će se po svojoj preciznosti u određivanju diskriminativnih čestica potrebnih za određivanje znanstvenog područja kojemu ispitanici pripadaju.

Metodologija

Metodologija - Uzorak

Odabir uzorka ovog istraživanja usko je vezan uz sadržaj i temu konstruiranog upitnika, kao i samu svrhu istraživanja te generalnu dostupnost ispitanika. Cilj istraživanja nameće nužnost korištenja studenata kao ispitanika ovog istraživanja, a pitanja, problemi i hipoteze ograničavaju populaciju na studente isključivo Sveučilišta u Zagrebu. Postavljanje takvog ograničenja uvedeno je iz više razloga, jedan od kojih je broj i podjela sastavnica unutar sveučilišta, koji nisu konzistentni na svim sveučilištima u Republici Hrvatskoj. Nekonzistentnost u strukturi sastavnica sveučilišta nepovoljno bi utjecala na grupiranje ispitanika pri analizi podataka. Drugi razlog za ograničavanje uzorka na studente Sveučilišta u Zagrebu jest dijeljena okolina. Iako se stavovi razvijaju od djetinjstva, a smatra se i da postoji genetski utjecaj na njihov razvoj, htjeli smo kontrolirati trenutnu lokaciju studija, odnosno prebivališta, radi otklanjanja utjecaja još jedne ometajuće varijable. Zbog dugoročnog utjecaja okoline koji postoji na razvoj stavova, ipak je uključeno pitanje o mjestu prebivališta u kojemu je ispitanik proveo najveći dio svog života.

Populacija kojoj se distribuirao upitnik jesu studenti svih sastavnica Sveučilišta u Zagrebu, a uzorak na kojemu su podaci prikupljeni te čiji su podaci analizirani predstavlja reprezentativan uzorak populacije te obuhvaća ispitanike sa sveukupno 22 različite sastavnice Sveučilišta u Zagrebu. Prema službenoj podjeli Sveučilišta po znanstvenim područjima, u trenutku određivanja ove klasifikacije, postoji podjela na *Prirodne, Tehničke, Biotehničke, Biomedicinske, Društvene* i *Humanističke* znanosti te područje *Umjetnosti*. Zastupljenost po znanstvenim područjima u našem uzorku je sljedeća: *Prirodne* - 3,21%, *Tehničke* - 28,45%, *Biotehničke* - 1,11%, *Biomedicinske* - 9,48%, *Društvene* - 30,26%, *Humanističke* - 20,22%, *Umjetnosti* - 0,7% te sastavnice kategorije *Ostalo* koje čine 6,57% uzorka.

Ukupni uzorak sastoji se od 717 ispitanika, od kojih 95,12% pripada dobnoj skupini 17-25 a preostalih 4,88% skupini 26-42. Ispitanici koji se identificiraju ženskim rodom čine 72,94% uzorka, dok se 25,66% identificira kao pripadnicima muškog roda. Preostalih 1,39% identificiraju se kao transrodne osobe, osobe nebinarnog rodnog identiteta i drugo.
Udio ispitanika koji su najveći dio svog života proveli u Zagrebu te mjestima u neposrednoj blizini (Velika Gorica, Dugo Selo i dr.) jest 44,77%, u odnosu na 55,23% ispitanika koji su veći dio života proveli na drugim područjima Republike Hrvatske.
Ciljna populacija ovog istraživanja jesu studenti Sveučilišta u Zagrebu, a iz prikupljenih demografskih podataka, možemo zaključiti da je uzorak reprezentativan za populaciju.

Metodologija - Instrument

U svrhu prikupljanja podataka za ovo je istraživanje konstruiran *Upitnik o navikama i stavovima studenata* (Dalje: *Upitnik*). *Upitnik* nije replikacija već postojećeg instrumenta, ali njegova konstrukcija djelomično je inspirirana *Personal Globe Inventory* mjernim instrumentom, koji mjeri različite aspekte interesa pojedinca. Veliki utjecaj pri konstrukciji imao je i Predigerov sferni model (1982) koji shematski prikazuje rezultate spomenutog inventara za različite aspekte posla, a odražava odnose između određenih vrsta poslova i interesa koje pojedinac iskazuje, prema rezultatima na mjernom instrumentu.
Upitnik o navikama i stavovima studenata online je upitnik, konstruiran na besplatnoj internetskoj platformi za kreiranje i distribuciju anketnih upitnika, platformi *Google Forms*. *Upitnik* se sastoji od uvoda i upute ispitaniku za rješavanje, skupa demografski orijentiranih čestica, 7 subskala, svaka od kojih označava jednu ispitivanu dimenziju te prostora za mogućnost odabira dobivanja povratne informacije o rezultatima istraživanja. Uvod i uputa ukratko informiraju ispitanike o svrsi prikupljanja podataka, o važnosti anonimnosti ispitanika, o etičkim odgovornostima istraživača te kontaktu istih. Prve čestice u *Upitniku* jesu *demografska pitanja*, a to su: godina rođenja, rod, mjesto odrastanja, fakultet, smjer te godina studija ispitanika. Nadalje, odvojene na zasebnim stranicama upitnika jesu subskale. Subskale i broj čestica koji pojedina skala sadrži su, redoslijedom kojim su bili konstruirani u *Upitnik*: *Glazba* (15), *Sport* (19), *Umjetnost i kultura* (16), *Ekologija* (11), *Tehnologija* (11), *Društvena pitanja i odgovornost* (12), *Vjerovanja* (8). Isključivši demografska pitanja te jedno pitanje u subskali *Sport* koje je podrazumijevalo *Da/Ne* odgovor, čestice u Upitniku bile su formulirane kao izjave, čiji je rezultat bio izražen Likertovom ljestvicom od 5 stupnjeva. Likertove ljestvice

pridružene česticama označavale su stupanj slaganja, odnosno stavove te čestinu, odnosno navike.

Stupnjevi na Likertovoj skali slaganja označavali su: (1) *Nimalo se ne slažem*, (2) *Ne slažem se*, (3) *Niti se slažem, niti se ne slažem*, (4) *Slažem se*, (5) *U potpunosti se slažem*. Stupnjevi na Likertovoj skali čestine označavali su: (1) *Nikada*, (2) *Rijetko*, (3) *Ponekad*, (4) *Često*, (5) *Vrlo često*.

Niže, u *Dodacima*, priloženi su primjeri čestica korišteni u *Upitniku*.

Upitnik o navikama i stavovima studenata konstruiran je s dva glavna cilja, pored onih koji uključuju zadovoljavanje metrijskih karakteristika testa, a tiču se odabira specifičnih dimenzija stavova koje su se ispitivale. Prvi je cilj bio da dimenzije budu dovoljno općenite i svakodnevne, kako bi svaki sudionik mogao i znao producirati valjan, ali i intuitivan odgovor. Drugi je cilj bio da dimenzije budu dovoljno zanimljive i dobno prilagođene kako bi privukle što veći broj ispitanika radi velike količine podataka koju zahtjeva eksploratorna analiza.

Metodologija - Postupak

Postupak prikupljanja podataka sačinjavaju dvije faze: faza predtestiranja i faza testiranja. Faza predtestiranja ima za cilj osigurati jednoznačni te što jednostavniji i gramatički točan upitnik kroz osvrte ispitanika koji pripadaju ispitivanoj populaciji. Predtestiranje se provodilo na 15 ispitanika, studenata Sveučilišta u Zagrebu, koji su ispunili prvu verziju *Upitnika*. Dodatna uputa koja im je dana ticala se davanja povratne informacije istraživačima o potencijalnim nejasnim rečeničnim strukturama, upitničkim česticama, gramatičkim pogreškama te okvirnom trajanju potrebnom za ispunjavanje upitnika. Od 15 prikupljenih osvrta, svi su bili pozitivno obojeni, s komentarima najčešće vezanim uz manje gramatičke pogreške te jednu česticu iz područja demografskih podataka koja je posljedično bila prilagođena sugestijama. Priroda osvrta nije bila negativna s obzirom na 92 testne čestice. Zbog prethodno iznesenih činjenica, u konačnoj analizi rezultata koristili su se rezultati predtestnih ispitanika te to nije smatrano narušavanjem metrijskih karakteristika rezultata.

Postupak prikupljanja podataka u testnoj fazi sastojao se od distribucije poveznice na online upitnik, na četiri različite razine: (1) osobno putem tekstualnih poruka (istraživač-potencijalni ispitanik, student Sveučilišta u Zagrebu), (2) putem objava u studijske i fakultetske grupe na društvenim mrežama, (3) putem objava na društvenim mrežama koje nisu bile ograničene na određene studijske ili fakultetske grupe, (4) putem službene e-mail molbe upućene

prodekanima za studente (i nastavu) određenih sastavnica Sveučilišta, radi daljnje distribucije upitnika.

Radi osiguravanja valjanosti rezultata, pristup rješavanju upitnika bio je ograničen, odnosno jedan je ispitanik mogao ispuniti *Upitnik* samo jednom, s iste e-mail adrese.

Budući da je distribucija, kao i rješavanje upitnika, bilo virtualno, sve upute o provedbi istraživanja te samom rješavanju bile su navedene u uputi na poveznici, prije početka rješavanja upitnika. Trajanje rješavanja upitnika bilo je 5-10 minuta.

Priprema i analiza podataka

Priprema i analiza podataka - Uzorak

Ispitanici iz uzorka pohađaju sveukupno 22 različite sastavnice Sveučilišta u Zagrebu koje smo grupirali u 3 veće grupe - *SOCIO* (društvene, humanističke i društveno-humanističke znanosti) - 56% uzorka, *STEM* (prirodne i tehničke znanosti i matematika) - 42,68% uzorka i *OSTALO* (Akademija likovne umjetnosti, Muzička akademija, Grafički fakultet te Tekstilno-tehnološki fakultet) - 1,26%, čija simbolička imena odgovaraju nazivu odgovarajućih klasa u programskom jeziku *Python* koji je korišten za analizu podataka. Grupi *STEM* pripadaju svi ispitanici koji pohađaju bilo koji od fakulteta: Fakultet elektrotehnike i računarstva, Fakultet strojarstva i brodogradnje, Fakultet kemijskog inženjerstva i tehnologije, Veterinarski fakultet, Prehrambeno-biološki fakultet, Farmaceutsko-biokemijski fakultet, Rudarsko-geološko-naftni fakultet, Medicinski fakultet, Stomatološki fakultet, Prirodoslovno-matematički fakultet, Građevinski fakultet te Fakultet organizacije i informatike. Grupi *SOCIO* pripadaju ispitanici koji pohađaju neki od fakulteta: Ekonomski fakultet, Fakultet političkih znanosti, Filozofski fakultet, Pravni fakultet te Učiteljski fakultet. Pri analizi skupa podataka metodama strojnog učenja bilo je potrebno eliminirati klasu *OSTALO* zbog nedovoljne frekvencijske značajnosti te činjenice da su neki modeli po svojoj prirodi binarni, stoga smo uz takve modele gledali ispitanike iz grupa *SOCIO* i *STEM* jer broje značajno više ispitanika. Također, za sve provedene testove nad grupom *OSTALO* pokazalo se da nisu statistički značajni, stoga možemo opravdano eliminirati upravo tu grupu kod modela strojnog učenja u svrhu primjene binarnih modela. Ovakav binarni model (s klasama *STEM* i *SOCIO* - studenti STEM područja te studenti društveno-humanističkih znanosti) odabrali smo jer nas je zanimala uspješnost klasifikacije tih konkretnih grupa studenata, umjesto analize modela s dvije klase od kojih se jedna sastoji od studenata koji studiraju isto područje, a druga od svih onih koji ne studiraju to područje (npr.

model s klasom *SOCIO* i *ne-SOCIO*). Intuitivno, u takvom modelu ne bi bilo potrebno eliminirati ispitanike klase *OSTALO* te bi koristili cijeli skup podataka.

Radi osiguravanja valjanosti rezultata, iz statističke analize izuzeti su oni ispitanici čiji su odgovori na demografska pitanja otvorenog tipa ukazivali na neprihvatljiv i neozbiljan pristup istraživanju. Takvih ispitanika bilo je dvoje, odnosno 0.28% od ukupnog uzorka.

Priprema i analiza podataka - Čestice

Kako bi pripremili skup podataka za modele strojnog učenja, moramo prvo obaviti statističku analizu podataka kako bi izolirali čestice koje pokazuju razliku u srednjoj vrijednosti, medijanu ili distribuciji s obzirom na klase *STEM* i *SOCIO*. Općenito, kod modela strojnog učenja želimo analizirati kako se ponaša model sa svim česticama koje opisuju neki objekt (tzv. *puni model* ili *puni skup*) naspram model s reduciranim brojem česticama, odnosno one koje nose najveću i najznačajniju količinu informacije o objektu kojeg opisuju. U našem slučaju, želimo reducirati skup od navedene 92 čestice na manji broj čestica koji podjednako dobro opisuje ispitanika, odnosno informaciju o tome u kojem području ispitanik studira. Kada dobijemo potencijalni minimalni skup čestica koje ulaze u model, ujedno dobivamo informaciju o tome na koje čestice su ispitanici najznačajnije različito odgovarali ovisno o području studiranja te možemo taj rezultat dalje interpretirati.

Prvi korak analize bio je da za svaku od 92 čestice usporedimo srednju vrijednost i medijan grupa *STEM* i *SOCIO*, te provedemo t-test o sredinama grupa čija je nulta hipoteza da su sredine dviju grupa jednake, a alternativa da grupe nemaju jednaku sredinu. Uz razinu značajnosti 0.001, reducirali smo skup od 92 čestice na 18 čestica (iz *Dodataka*: 2.1., 2.3., 2.4., 2.9., 3.8., 3.10., 3.13., 3.14., 3.15., 5.1., 5.3., 5.7., 5.9., 5.10., 5.11., 7.1., 7.2., 7.3.) čija je p-vrijednost testa iznosila <0.001. Od ukupno 18 čestica 4 dolaze iz dimenzije *Sport*, 5 iz dimenzije *Umjetnost i kultura*, 5 iz dimenzije *Tehnologija* te 3 iz dimenzija *Vjerovanja*. Statističkom analizom, sukladno, utvrdili smo da nema značajne razlike u odgovorima ispitanika grupe *STEM* i *SOCIO* niti na jedno pitanje dimenzije *Glazba*, *Ekologija* i *Društvena odgovornost*. Demografski podaci također ne pokazuju značajnu korelaciju s područjem studiranja, stoga oni neće ulaziti u modele strojnog učenja.

Drugi dio redukcije broja čestica koje ulaze u konačni model napravit ćemo pomoću Logističke regresije (Walpole, Myers, Myers i Ye, 1993), statističkog modela koji se također koristi u strojnom učenju. Pomoću programskog jezika *Python* možemo dati modelu Logističke regresije 18 čestica koje smatramo statistički najznačajnijim. Za njih postoji prethodno

implementirana funkcija koja ispisuje sažetak modela u kojem, između ostalog, možemo saznati koji regresori, odnosno čestice, su značajni, a koji nisu. Naime, kod modela logističke regresije možemo provoditi testove nad regresorima (koji su kod nas čestice) te ćemo se osloniti upravo na test koji testira nultu hipotezu: *Regresor (čestica) nije značajan* naspram alternative: *Regresor je značajan*. Na razini značajnosti 0.10 sažetak Logističke regresije s 18 čestica pokazuje da je samo 8 čestica značajno u modelu. Regresori koji su pokazali da nisu značajni također pokazuju visoke koeficijente korelacije sa jednim od 8 značajnih koeficijenata, te ih iz tog razloga možemo ukloniti pri redukciji broja čestica.

Čestice koje nose najveću količinu informacija o području studiranja ispitanika su, redom, u Dodacima: 2.2., 2.10., 3.9., 3.11., 3.15., 5.3., 5.7., 7.1. Za detaljniji pregled čestica, osvrnite se na *Čestice* priložene u *Dodacima*.

Navedene skupove od 92 čestice (puni skup), 18 čestica (prvi reducirani skup) i 8 čestica (minimalni skup) analizirat ćemo nad sljedećim modelima nadziranog strojnog učenja: Naivni Bayesov klasifikator (Leung, 2007), Stablo odluke (Charbuty i Abdulazeez, 2021) i Random Forest algoritam (Burkov, 2019), Strojeve s vektorom potpore (Suthaharan, 2016) te Logističku regresiju.

Kada pričamo o nadziranom strojnom učenju, govorimo o praksi gledanja skupa podataka kao niz *označenih primjera* $(x_i, y_i)_{N=i}$ pri čemu x_i nazivamo vektorom značajki, a značajke su brojevi, riječi ili neke kompleksnije strukture koje na neki način opisuju izlaznu varijablu, odnosno y_i (Burkov, 2019) dok *N* predstavlja broj označenih primjera u skupu podataka. Dimenzija vektora značajki jednaka je broju značajki koji opisuje neku izlaznu varijablu. Za primjer jednostavnog označenog primjera, izlazna varijabla može biti informacija o tome ima li osoba srčanu manu (ona poprima vrijednost *DA* ukoliko osoba ima srčanu manu te *NE* ukoliko nema), a 3-dimenzionalni vektor značajki može biti vektor koji sadrži 3 informacije: spol osobe, njihovu dob i tjelesnu masu. U našem skupu podataka, vektor značajki sadržavat će odgovore ispitanika na određenih *n* čestica (n = 92 za puni skup, 18 za prvi reducirani skup te 8 za minimalni skup) koji zajedno opisuju izlaznu varijablu - informaciju o području studiranja ispitanika (klasa *SOCIO* ili *STEM*). Kod nadziranog strojnog učenja, niz označenih primjera dijelimo na dva skupa - skup za treniranje i skup za testiranje. U skup za treniranje u praksi ulazi 70-80% cijelog niza (dok preostali označeni primjeri pripadaju skupu za testiranje), te njega koristimo kako bi "istrenirali" ili prilagodili model. Model se "uči" na skupu za treniranje, te dobiveni model testiramo na novim ili neviđenim podacima, odnosno skupu za testiranje. Ideja je da za svaki primjer iz skupa za testiranje modelu predamo vektor značajki te da nam model da svoju predikciju izlazne varijable, koju ćemo nakon toga usporediti sa

stvarnom izlaznom varijablom tog primjera. Udio točno predviđenih primjera u skupu za treniranje zovemo *preciznošću* modela, te rangiramo modele prema uspješnosti upravo s obzirom na preciznost.

Osim preciznosti, za usporedbu različitih modela također koristimo metodu unakrsne validacije. Unakrsna validacija zasniva se na podjeli skupa podataka na blokove pri podjeli na skup za treniranje i skup za testiranje (Refaeilzadeh, Tang i Liu, 2009). Za primjer, ako odredimo da je veličina skupa za treniranje 20% cijelog niza označenih primjera, to znači da bi cijeli skup podataka mogli podijeliti na pet blokova tako da se u svakom bloku nalazi 20% podataka. Ako na istom skupu podataka primijenimo dva puta isti model, ali uzmemo različite blokove kao skup za testiranje, dobit ćemo različite rezultate s obzirom na prilagodbu modela (tada će i skupovi za treniranje biti različiti). Ovdje dolazi metoda unakrsne validacije, koja navedeni "problem" rješava analizirajući sve moguće kombinacije blokova te daje procjenu prilagodbe bazirane na svim modelima (npr. ukoliko je preciznost modela 75%, predviđana preciznost metodom unakrsne validacije može biti $76 \pm 4\%$).

Rezultati

Puni skup od 92 čestice, prvi reducirani skup od 18 čestica te minimalni skup od 8 čestica ispitali smo na spomenutim modelima te su rezultati preciznosti i predviđane preciznosti dobivenom metodom unakrsne validacije. Na dolje prikazanoj tablici u ćelijama se nalazi postotak preciznosti u prvom, a postotak preciznosti utvrđene metodom unakrsne validacije u drugom retku iste ćelije.

Tablica 1. Preciznost predviđanja klase na neviđenim primjerima kod modela nadziranog strojnog učenja.

	Originalni skup (92 čestice)	Prvi reducirani skup (18 čestica)	Minimalni skup (8 čestica)
Naivni Bayesov klasifikator	70.21% 71 ± 5%	73.05% 75 ± 11%	78.01% 76 ± 6%
Stablo odluke	71.63% 68 ± 14%	61.7% 69 ± 11%	71.63% 68 ± 7%
Random Forest	73.76% 72 ± 11%	76.6% 75 ± 10%	74.47% 75 ± 3%
Logistička regresija	75.18% 74 ± 8%	76.6% 79 ± 6%	77.31% 79 ± 8%
Stroj s vektorom potpore	75.89% 79 ± 8%	74.47% 78 ± 6%	78.01% 78 ± 7%

Kod modela logističke regresije imamo dodatnu mjeru kvalitete prilagodbe, a to je koeficijent pseudo-R^2. Naime, R^2 ili koeficijent determinacije mjera je koja kod linearne regresije predstavlja omjer sume kvadratnih reziduala i zbroja sume kvadratnih reziduala s kvadratnim greškama, a njezin iznos poprima vrijednosti u intervalu [0, 1]. Kod logističke regresije, kakogod, nemamo reziduale kao takve jer se radi o binarnoj klasifikaciji (klasa 0-*SOCIO* i 1-*STEM*), no postoji pseudo-koeficijent determinacije ili pseudo-R^2, čija je interpretacija ekvivalentna koeficijentu kod linearne regresije - veći pseudo-R^2 označava bolju prilagodbu. Pseudo-R^2 punog skupa od 92 značajke iznosi 0.455, 0.326 kod prvog reduciranog skupa te 0.306 kod minimalnog skupa.

Što se tiče 8 konačnih čestica koje su bile korištene u analizi, one su, po silaznoj statističkoj značajnosti u regresijskom modelu navedene niže, kao i smjer zastupljenosti odgovora pojedine grupe u svakoj od čestica. Osam čestica, kao i grupa koja je u prosjeku s većom zastupljenosti odgovarala u korist tvrdnji, jesu:

- *Moje znanje programiranja natprosječno je s obzirom na prosječnog građanina RH* (prosječna ocjena je veća u grupi *STEM*),
- *U slobodno vrijeme čitam knjige koje nisu stručna literatura* (prosječna ocjena je veća u grupi *SOCIO*),
- *Vjerujem u horoskop* (prosječna ocjena je veća u grupi *SOCIO*),
- *Pratim nogomet* (prosječna ocjena je veća u grupi *STEM*),
- *U slobodno vrijeme aktivno pratim sport* (prosječna ocjena je veća u grupi *STEM*),
- *U slobodno vrijeme posjećujem muzeje* (prosječna ocjena je veća u grupi *SOCIO*),
- *U slobodno vrijeme posjećujem kazalište* (prosječna ocjena je veća u grupi *SOCIO*)
- *Volio/voljela bih da sam kroz sustav osnovnoškolskog obrazovanja imao/la više mogućnosti upoznati se s predmetima iz područja elektrotehnike i računarstva* (prosječna ocjena je veća u grupi *SOCIO*)

Rasprava

Općenito preciznosti modela poprimaju vrijednosti u intervalu [61.7%, 78.01%] te najmanju preciznost pokazuje skup od 18 čestica na modelu Stabla odluke, a najveću skup od 8 čestica na modelu Stroja s vektorom potpore. Ovaj rezultat je očekivan, jer je model Stabla odluke podložan prenaučenosti ili pojavi prekomjerne prilagodbe modela skupu za treniranje, tako da kada model primi nove, neviđene podatke pokazuje znatno manju preciznost predviđanja izlazne varijable. Još jedan razlog zašto stablo odluke pokazuje lošije performanse u odnosu

na ostale modele jest da ono ne diskriminira čestice po značajnosti (kao što Linearna/Logistička regresija i Strojevi s vektorom potpore to rade) već se oslanja na entropiju - uređenost skupa. U tablici je primjetno da se redukcijom skupa s 92 na 8 čestica kod svih modela osim stabla odluke povećava preciznost predviđanja izlaza na neviđenim podacima (kod stabla odluke preciznost ostaje ista). Uzevši u obzir unakrsnu validaciju, dva najuspješnija modela su Logistička regresija, Stroj s vektorom potpore (SVM) te Naivni Bayes, koji pokazuju preciznosti 77.31% (Logistička regresija), odnosno 78.01% (SVM i Naivni Bayes) . Ovaj rezultat je očekivan jer oba modela diskriminiraju čestice s obzirom na značajnost, odnosno one čestice za koje algoritam procijeni da nose zanemarivu količinu informacije o izlaznoj varijabli neće podariti jednaku težinu ili važnost pri treniranju modela. Kod Logističke regresije tu pojavu možemo i primijetiti kod koeficijenata regresora (značajniji regresori uz sebe koeficijent veće apsolutne vrijednosti).

Pojava povećanja preciznosti kada smanjimo broj čestica koje predajemo modelu možemo objasniti upravo time da smo uspješno reducirali skup čestica, odnosno eliminirali one čestice koje ne pridonose predviđanju izlazne varijable. Laički, umjesto toga da imamo model kojemu "kažemo" 92 podatka o našem ispitaniku, a model na osnovu tih podataka predviđa u kojem području ispitanik s određenom preciznošću, vidimo da ukoliko modelu damo 8 najrelevantnijih informacija o ispitaniku, on će predvidjeti njegovo područje studiranja s nešto većom preciznošću. To podrazumijeva da čestice na koje su ispitanici obje grupe vrlo slično odgovarali (t-testom nismo mogli odbaciti hipotezu da su sredine grupa jednake, medijani su im jednaki i distribucije slične) te sukladno ne opisuju dobro ispitanika s obzirom na područje studiranja - stvaraju određen "šum" u nizu označenih primjera, odnosno povećavaju entropiju (količinu neuređenosti) skupa. Kada njih uklonimo, u očima algoritma izbacili smo informacije koje nisu relevantne te algoritmi strojnog učenja predviđaju izlaznu varijablu novim podacima s većom preciznošću. Praksa uklanjanja takvih čestica česta je kod modela strojnog učenja radi postizanja generalizacije, koja postiže veću preciznost i sprječava spomenutu prenaučenost (Neyshabur, Bhojanapalli, McAllester i Srebro, 2017).

Navedeno "uklanjanje šuma" pomoću redukcije skupa čestica također možemo primijetiti kod metode unakrsne validacije. Podsjetimo se, metodom unakrsne validacije dajemo procjenu stvarne preciznosti predviđanja modela. Ona je dana u obliku $X \pm 2 * SD\ \%$, pri čemu X predstavlja procjenu preciznosti, a SD označava standardnu devijaciju procjene. Vidimo u tablici da se standardna devijacija smanjuje kako smanjujemo skup čestica, te nam je to dodatna potvrda da smo eliminirali čestice koje nisu pridonosile predviđanju izlazne varijable.

Konačno, možemo analizirati minimalni skup i primijetiti da modeli pokazuju relativno visoke preciznosti predviđanja izlazne varijable te da je najveća zabilježena preciznost 78.01%, što znači da su ti modeli (konkretno model Naivni Bayes i Stroj s vektorom potpore) u 78.01% slučajeva točno klasificirali ispitanika u područje *STEM* ili *SOCIO*.

Detaljnom kvalitativnom analizom čestica prema kojima smo odredili klasifikaciju studenata na dvije, prethodno spomenute grupe, mogu se ustanoviti dvije stvari. Prvo, jasno je da ne postoje razlike u svim ispitivanim dimenzijama, već samo u rijetkim česticama određenih dimenzija. Dimenzije u kojima smo pronašle diskriminativne čestice jesu *Tehnologija* i *Umjetnost i kultura* te *Sport* i *Vjerovanja*. Takvi rezultati ukazuju na to da se studenti, generalno, ne razlikuju po mnogočemu, na razini populacije te da se njihovi stavovi i navike distribuiraju po normalnoj distribuciji. razlike koje jesu pronađene i po kojima se studente može diferencirati i klasificirati jesu u temama koje su specifične za njihove struke, ili vezane uz interese koje bismo pretpostavili da bi studenti određenog znanstvenog područja mogli imati.

Ishod naše pripreme podataka, kao i njihova konačna analiza, ukazuju na to da uistinu postoje pretpostavljene razlike među studentima, a koje imaju temelje u znanstvenom području kojemu oni pripadaju. Iako ne iz mnogobrojnih, te je razlike moguće zabilježiti iz 8 empirijski *značajnih* čestica prema kojima su modeli Logističke regresije i Stroja s vektorom potpore studente klasificirali u njihova pripadajuća znanstvena područja, sa značajnim uspjehom. Korisnost modela u odgovaranju na postavljena istraživačka pitanja te mogućnost ustanovljenja razlike među studentima na temelju znanstvenih područja kojima pripadaju, potvrđuje i istinitost druge hipoteze, odnosno pretpostavke o korisnosti modela nadziranog strojnog učenja kod obrade opisanog seta podataka i utvrđenog cilja. Konačno, razlike u preciznostima modela, koje se očituju u broju čestica korištenih pri analizi, potvrđuju istinitost treće hipoteze koja pretpostavlja spomenute razlike između modela korištenih u analiza rezultata ovog istraživanja.

Zaključno, analizom i interpretacijom dobivenih rezultata, opisujemo sve tri predložene hipoteze istinitima te empirijski potvrđenima.

Zaključak

Kao pitanje na koje smo htjeli pronaći odgovor u ovome istraživanju imali smo sposobnost razlikovanja studenata Sveučilišta u Zagrebu korištenjem metoda strojnog učenja, na temelju stavova o njima dobro poznatim temama te navikama koje uz njih vežemo. Također, željeli

smo provjeriti korisnost modela strojnog učenja u analizi seta podataka kao što je ovaj s kojime raspolažemo. Osim što smo nakon provedbe istraživanja proveli detaljnu, višestruku analizu četirima metodama nadziranog strojnog učenja, ona se pokazala i kao korisnom metodom, odnosno alatom za odgovaranje na pitanja postavljena u ovom istraživanju. Također, pokazalo se kako, iako ne mnogobrojne, postoje diskriminatorne točke, odnosno čestice prema kojima grupe studenata možemo razlikovati. Spomenute čestice mogu se koristiti kako bi podijelile studente na dvije, sada više ne samo arbitrarne i proizvoljne, već i empirijski različite skupine, a to su *STEM* i *SOCIO*. Skupine, za koje se i laički smatra da su poprilično različite, razlikovale su se u česticama na dimenzijama u kojima bismo razlike očekivali, s obzirom na znanstvena područja koja skupine obuhvaćaju. Razlike među skupinama *STEM* i *SOCIO* postojale su i u česticama na dimenzijama koje usko ne možemo povezati s potrebama struka uključenima u dvije skupine. Unatoč rezultatima koji ukazuju na razlike u navikama, više nego stavovima, među dvije skupine zagrebačkih studenata, važno je naglasiti kako ni neizrečen cilj, niti želja ovog istraživanja nije povećanje predrasuda i stereotipa u društvu, pogotovo među populacijom koja je istima svakodnevno opterećena. Umjesto za poticanje razdjela i traženja razloga za grupiranjem, rezultati ovog istraživanja mogu biti gledani u pozitivnom svjetlu, na dva načina: kao uspjeh s obzirom na relativno visoke preciznosti modela strojnog učenja te kao potencijalnu praktičnu iskoristivost u grani organizacijske psihologije. Poznato je kako dijeljene vrijednosti pojedinca i kompanije, odnosno poslodavca, visoko pozitivno koreliraju pri odabiru mjesta zapošljavanja. Provodeći istraživanja kao što je ovo, koja se bave stavovima, navikama, ali i ona koja se bave interesima, mišljenjima i sličnim konstruktima, možemo doprinijeti boljoj informiranosti o populaciji koja stupa na tržište. Takvi podaci bili bi korisni raznim kompanijama i marketinškim agencijama radi optimizacije selekcijskog procesa i inicijalnog privlačenja ciljane populacije, ali i cjelokupnog poboljšanja radnog okruženja.

Dodaci

Čestice iz Upitnika o navikama i stavovima studenata, prema ispitivanim dimenzijama: (1) Glazba, (2) Sport, (3) Umjetnost i kultura, (4) Ekologija, (5) Tehnologija, (6) Društvena pitanja i odgovornost, (7) Vjerovanja.

Čestice skale slaganja:

1.1. Uživam u slušanju glazbe.

1.2. Uživam u izlascima u noćne klubove.

1.3. Izloženost određenim žanrovima glazbe kod nekih ljudi potiče agresivno ponašanje.

1.4. Predrasude koje postoje oko ljubitelja određenih glazbenih žanrova uglavnom su točne.

Čestice skale čestine:

1.5. U slobodno vrijeme slušan glazbu.

1.6. Slušam glazbu koja se pojavljuje na „Top 100“ i sličnim ljestvicama.

1.7. Prije COVID-19 pandemije, posjećivao/la sam noćne klubove.

1.8. Rock glazbu slušam…

1.9. Turbofolk glazbu slušam…

1.10. Jazz glazbu slušam…

1.11. Metal glazbu slušam…

1.12. Hrvatske hitove/treš glazbu slušam…

1.13. Hip-hop/rap glazbu slušam…

1.14. Indie glazbu slušam…

1.15. Elektroničku glazbu slušam…

Čestica koja zahtjeva Da/Ne odgovor:

2.1. Do sada u životu sam se natjecateljski bavio/la, ili se još uvijek bavim nekim sportom.

Čestice skale slaganja:

2.2. U slobodno vrijeme aktivno pratim sport, sportske vijesti, i/ili sportske rezultate.

2.3. Kada pratim sport, to je obično tijekom većih sportskih događanja kao što su Olimpijske igre, Svjetska prvenstva i sl.

2.4. Dok gledam sportske utakmice, mečeve i dr., osjećam se uzbuđeno i emotivno uključeno.

2.5. Uživam u gledanju sporta ili komentiranju sportskih rezultata s prijateljima.

2.6. Dok gledam sportske utakmice, mečeve i dr., teško mi je uživjeti se u natjecateljsku atmosferu.

2.7. Smatram da je gledanje sporta „gubitak vremena“.

2.8. Djecu bi trebalo poticati na bavljenje sportom.

Čestice skale čestine:

2.9. Rekreativno se bavim sportom ili nekom drugom fizičkom aktivnošću.

2.10. Pratim nogomet.

2.11. Pratim tenis.

2.12. Pratim boks.

2.13. Pratim rukomet.

2.14. Pratim Formulu 1.

2.15. Pratim košarku.

2.16. Pratim MMA.

2.17. Pratim atletiku.

2.18. Pratim vaterpolo.

Čestice skale slaganja:

3.1. Slabo se razumijem u umjetnost.

3.2. Umjetnost je temeljni dio svake kulture.

3.3. U osnovnoj školi bili smo premalo izlagani umjetnosti.

3.4. Kada netko spomene umjetnost, moja prva asocijacija je slikarstvo.

3.5. Kazališne predstave zastarjeli su oblik filma.

3.6. Današnje društvo podcjenjuje moć literarnog izražavanja.

3.7. Sustav osnovnoškolskog i srednješkolskog obrazovanja naučio me uživati u čitanju.

Čestice skale čestine:

3.8. Posjećujem kino.

3.9. Posjećujem kazališta.

3.10. Posjećujem koncerte klasične glazbe.

3.11. Posjećujem muzeje i/ili galerije.

3.12. Posjećujem nastupe stand-up komedije.

3.13. Teško mi je uživjeti se u umjetnička djela.

3.14. Muzeje posjećujem samoinicijativno, iz užitka.

3.15. U slobodno vrijeme čitam knjige koje nisu stručna literatura.
3.16. Muzeji su mi dosadni.

Čestice skale slaganja:
4.1. Znam definirati područja koja proučava znanost ekologije.
4.2. Svjestan sam negativnih utjecaja koje donose klimatske promjene.
4.3. Brine me problem globalnog zatopljenja.
4.4. Nisu mi dostupne informacije o uzrocima i/ili načinima borbe protiv klimatskih promjena.
4.5. Smatram da kao pojedinac ne mogu pridonijeti značajnoj promjeni u borbi protiv klimatskih promjena.
4.6. Ciljevi i direktive koje EU postavlja u borbi protiv klimatskih promjena samo su formalnosti i „prazne riječi".
4.7. Velike su kompanije spremne prilagoditi način rada kako bi pomogle u borbi protiv klimatskih promjena, čak i ako ta prilagodba dovodi do gubitka profita.
4.8. Savjesno i obzirno koristim energiju u svome kućanstvu.
Čestice skale čestine:
4.9. Kada kupujem odjeću, kupujem u prodavaonicama „brze mode" (ZARA, H&M, Pull&Bear, i dr.).
4.10. U svome kućanstvu razvrstavam otpad.
4.11. Koristim plastične proizvode za jednokratnu uporabu.

Čestice skale slaganja:
5.1. Aktivno pratim vijesti o tehnološkim proizvodima i novitetima koji se plasiraju na tržište.
5.2. Briga oko gubitka posla zbog robotizacija opravdana je briga ljudi današnjice.
5.3. Moje znanje iz programiranja je iznadprosječno s obzirom na stanovništvo Republike Hrvatske.
5.4. Smatram da kriptovalute imaju značajnu ulogu na današnjem tržištu.
5.5. Smatram da ulaganje u kriptovalute dugoročno nije isplativo.
5.6. Smatram da je ulaganje u NFT-ove dugoročno isplativo.
5.7. Volio/voljela bih da sam kroz sustav osnovnoškolskog obrazovanja imao/la više mogućnosti upoznati se s predmetima iz područja elektrotehnike i računarstva.
Čestice skale čestine:
5.8. Koristim računalo za fakultetske i/ili poslovne obveze.

5.9. U slobodno vrijeme bavim se programiranjem.
5.10. U slobodno vrijeme igram videoigre.
5.11. Koristim YouTube, Twitch i slične platforme kako bih gledao druge ljudi kako igraju videoigre.

Čestice skale slaganja:
6.1. O politici bi trebali govoriti samo oni koji se njome bave.
6.2. Za svako pravo koje imam postoji i popratna, njemu sukladna odgovornost.
6.3. Stanovnici demokratskih republika koji imaju pravo glasa trebali bi to pravo aktivno i koristiti.
6.4. Stalo mi je do toga tko će upravljati mojim gradom, mojom županijom i državom.
6.5. Vidim budućnost u kojoj hrvatska politika djeluje bez korupcije i kriminala.
6.6. Prosvjedi su efektivan način postizanja cilja.
6.7. Namjera ljudi koji organiziraju prosvjede aktivno je izazivanje nemira i sukoba.
6.8. Društvene mreže su prekrcane informacijama o raznim oblicima aktivizma.
6.9. Većinu objava na društvenim mrežama koje se tiču aktivizma smatram iskrenim pokušajima promjene u društvu.
6.10. Aktivizam na društvenim mrežama postaje mi zamoran.
6.11. Znam gdje mogu pronaći provjerene informacije o društvenom pitanju koje me zanima ili brine.
Čestice skale čestine:
6.12. Kao glasač sudjelujem u izborima u Republici Hrvatskoj.

Čestice skale slaganja:
7.1. Vjerujem u horoskop.
7.2. Upoznat/a sam sa svojom natalnom kartom.
7.3. Vjerujem u sudbinu – u predodređenost stvari i događaja.
7.4. Vjerujem u teoriju evolucije.
7.5. Vjerujem u postojanje nekog oblika života poslije smrti.
7.6. Vjerujem u postojanje izvanzemaljskih civilizacija.
7.7. Sklon/a sam istraživanju raznih teorija zavjere.
7.8. Svijet bi se jasno mogao podijeliti na „dobro“ i „zlo“.

Literatura

Akert, R. M., Aronson, E., Wilson, T. D. (2005). *Uvod u socijalnu psihologiju.* Zagreb: Mate d.o.o.

Burkov, A. (2019). *The hundred-page machine learning book*, *Volume (1)*. Quebec City, QC, Canada: Andriy Burkov.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2, Volume(1)*, 20-28.

Leung, K. M. (2007). Naive Bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 123-156.

Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet*, *383*(9921), 999-1008.

Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. *Advances in neural information processing systems*„ 1-3.

Obaid, O. I., Mohammed, M. A., Ghani, M. K. A., Mostafa, A., & Taha, F. (2018). Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. *International Journal of Engineering & Technology*, *7*(4.36), 160-166.

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interest and occupations? *Journal of Vocational Behavior, 21*(1), 259–287.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, *5*, 532-538.

Subasi, A. (2020). *Practical Machine Learning for Data Analysis Using Python*. Academic Press.

Suthaharan, S. (2016). Support vector machine. In *Machine learning models and algorithms for big data classification*, Springer, Boston, MA.

Šverko, I. (2008). Spherical model of interests in Croatia. *Journal of Vocational Behavior, 72*(1), 14-24.

Tesser, A. (1993). The Importance of Heritability in Psychological Research: The Case of Attitudes. *Psychological Review, 100*(1), 129-142.

Tracey, T. J. G. (2021). *PGI. Personal Globe Inventory [Test description, manual, questionnaires PGI-Activities, PGI-Occupations, PGI-Short, PGI-Mini, scoring PGI/PGI-Short/PGI-Mini, and case examples]*. In Leibniz Institute for Psychology (ZPID) (Ur.), Open Test Archive. https://doi.org/10.23668/psycharchives.4545

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (1993). *Probability and statistics for engineers and scientists*, *Volume (5)*. New York: Macmillan.

Sažetak

Klasifikacija studenata prema znanstvenom području kojem pripadaju na temelju iskazanih stavova i navika metodama strojnog učenja

Autorice ovog rada su Barbara Geld, studentica Psihologije na Filozofskom fakultetu i Rea Pintar, studentica Računarstva na Fakultetu elektrotehnike i računarstva. Ideja ovog interdisciplinarnog istraživačkog rada leži u metodologiji psihološkog istraživanja te izradi upitnika u sklopu takvog istraživanja, kao i analiza podataka izvršena u okviru paradigmi postavljenih u području računarstva te računalnih znanosti. Tema istraživanja bila je istražiti razlike u navikama i stavovima studenata Sveučilišta u Zagrebu. Cilj nam je također bio ispitati možemo li na temelju prikupljenih odgovora na konstruirani upitnik klasificirati studente prema njihovom području studiranja. Koristeći upitnik kreiran konkretno za ovo istraživanje, prikupili smo podatke o navikama i stavovima studenata u određenim aspektima svakodnevnog života poput glazbe, sporta, korištenja tehnologije i sl. Analiza podataka, koja je podrazumijevala određene metode iz područja strojnog učenja potvrdila je da su predložene hipoteze s kojima smo ustupili u ovo istraživanje doista istinite. Studente možemo, prema odgovorima upitnika, s relativno visokom preciznošću klasificirati u grupe prema njihovom

području studiranja te su se metode strojnog učenja pokazale efikasnim i korisnim u provedenoj analizi.
Ključne riječi: Stavovi; Navike; Interdisciplinarnost; Strojno učenje

Abstract

Classification of university students by field of study based on their habits and attitudes using Machine learning methods

The authors of this paper are Barbara Geld, Psychology student at the Faculty of Humanities and Social Sciences and Rea Pintar, Computer Science student at the Faculty of Electrical Engineering and Computing. The basis of this interdisciplinary research paper lies in psychological research methodology and questionnaire construction, as well as the data analysis done within the paradigms set in computational sciences. The topic of research was to explore the potential differences in attitudes and habits between students of the University of Zagreb. Our goal was also to explore whether or not it is possible to create classifications from results acquired, based on the scientific field they are doing their studies in. By using an originally constructed questionnaire, we have collected data on students' attitudes and habits in certain areas of everyday life - such as Music, Sport, Ecology, Technology, etc. The analysis, which consisted of multiple methods within the machine learning system, had concluded that the hypotheses proposed in the beginning of the research are indeed empirically proven to be true. The students can be classified into groups by their field of study, based on the results of the questionnaire, and the models of the machine learning system had been a useful tool in said analysis.
Key words: Attitudes; Habits; Interdisciplinarity; Machine Learning