Interpretabilno strojno učenje

Dejan Ljubobratović
Odjel za informatiku, Sveučilište u Rijeci, Rijeka, Hrvatska

I. SAŽETAK

Interpretabilnost modela strojnog učenja u zadnje je vrijeme jedna od važnijih tema na području strojnog učenja. Ona predstavlja uvjet primjene strojnog učenja u domenama u kojima ekspert s područja primjene osim što u raznim fazama analize podataka ugrađuje svoju ekspertizu čime usmjerava cijeli proces učenja modela dodatno i evaluira konačne rezultate koje je iz tog razloga potrebno predstaviti u čovjeku razumljivom obliku. U radu se predstavlja širi kontekst problematike interpretabilnog strojnog učenja. Na dva konkretna primjera interpretabilnog strojnog učenja daje se ilustracija primjene metoda iz ovog područja u različitim domenama. Prvi primjer primjene odabran je u području dubinske analize podataka s područja edukacije, a drugi je primjer iz područja inteligentnog upravljanja u hladnom lancu voća.

Još od 2010. godine Sveučilište u Rijeci koristi LMS sustav baziran na Moodle platformi kojim se nadopunjava tradicionalni način poučavanja. LMS se koristi za dijeljenje dokumenata, ispite, kvizove, video lekcije, praćenje uspjeha studenata i mnogo više. Svaki put kada student, koristeći svoj online račun, pristupi LMS sustavu kreira se log zapis u kojega se pohranjuju njegove aktivnosti. Ti zapisi su upotrijebljeni za kreiranje skupa podataka od nekoliko stotina opservacija.

U ovom radu korišten je algoritam slučajnih šuma (engl. Random forest algorithm) u svrhu predikcije prolaznosti studenata (grade) koristeći prediktore (lectures, quizzes, labs i videos) koji su dobiveni iz logova sustava za upravljanje učenjem (engl. Learning management system, LMS) Moodle.

Primjenu algoritma slučajnih šuma i postupaka interpretacije modela ilustriramo u uvodnim razmišljanjima i početnim rezultatima o problemu zaključivanja o zrelosti voća na temelju izmjerenih senzorskih podataka i pozadinskog znanja eksperta iz domene primjene.

Izrada modela predviđanja koristeći algoritam slučajnih šuma je relativno jednostavna ako ju usporedimo sa interpretacijom dobivenih rezultata. Interpretacija modela slučajnih šuma kao i svih ostalih „black box“ modela predstavlja izazov s obzirom na složenost njihovih mehanizama za donošenje odluka.

Postoje brojne nove tehnike koje nam pomažu u tom poslu, a nekolicina njih je predstavljena u ovom radu.

Izazov s kojim se također susreću mnogi istraživači prilikom korištenja „black box“ algoritama je Europska Opća uredba o zaštiti podataka (GDPR). GDPR ima značajan utjecaj na mnoge aspekte prikupljanja i obrade podataka građana EU.

U ovom radu istaknuti ćemo najvažnija ograničenja GDPR-a na dubinsku analizu podataka, uključujući „pravo na objašnjenje“.

Ključne riječi: LMS sustav, algoritam slučajnih šuma, dubinska analiza podataka s područja edukacije, predviđanje uspjeha studenata, interpretabilnost, interpretabilno strojno učenje

II. UVOD

Interpretabilno strojno učenje moglo bi se definirati kao uporaba modela strojnog učenja u svrhu otkrivanja relevantnog znanja o odnosima domena sadržanih u podacima. Ovdje znanje smatramo relevantnim samo ako nam može pružiti uvid u odabrani problem domene (Murdoch et al. 2019).

Neke algoritme dubokog učenja, takozvane "black box" algoritme, gotovo je nemoguće protumačiti. Njihova složenost nam uvelike otežava da razumijemo zašto i kako je model strojnog učenja donio određenu odluku (Ribeiro, Singh, and Guestrin 2016). Na primjer, ako banka osobi odbije kredit "black box" algoritmom, bi li ta osoba trebala znati koji su čimbenici pridonijeli toj odluci? Ima li pravo znati zašto je odbijena i imati priliku žaliti se. To je vrlo teško saznati ako je odluka donesena "black box" algoritmom. S druge strane, davanje dovoljnog broja objašnjenja moglo bi omogućiti reverzni inženjering procesa odlučivanja i tako omogućiti potencijalnim kriminalcima da prevare sustav i promijene ishod (Piatetsky-Shapiro 2018).

U prvom primjeru postupke interpretabilnog strojnog učenja (Interpretable Machine Learning, IML) primjenjujemo u području dubinske analize podataka u domeni edukacije.

Dubinska analiza podataka i tumačenje podataka prikupljenih na Massive Online Open tečajevima (MOOC) dobro je istraženo i popularno, te pruža istraživaču ogromnu bazu različitih zapisa (Nagrecha, Dillon, and Chawla 2019). Na primjer, samo je jedan tečaj “Learning How to Learn: Powerful mental tools to help you master tough subjects” koji je ponudilo Sveučilište McMaster (University of California, San Diego) upisalo više od 1,7 milijun ljudi pomoću platforme Coursera (“Learning How to Learn: Powerful Mental Tools to Help You Master Tough Subjects” 2019), dok LMS sustavi poput Moodle-a koji se koriste za dopunu obrazovnih procesa na sveučilištima i u školama barataju sa znatno manjom količinom podataka. U ovom istraživanju izgraditi ćemo model za predviđanje uspjeha studenata (*grade*) kao funkciju aktivnosti na kolegiju koristeći algoritam slučajnih šuma.

U ovom radu koristili smo nekoliko metoda za interpretaciju zadanog modela dajući objašnjenja rezultatima algoritma slučajnih šuma. Za istraživanje podataka, model predviđanja i za interpretaciju rezultata u ovom radu koristili smo R jezik v. 3.6.1.

Rad je podijeljen u četiri logička dijela. U prvom dijelu su predstavljene osnovne ideje pronađene u sličnim istraživanjima, nakon čega smo istaknuli najizazovnija GDPR ograničenja u dubinskoj analizi podataka.

Drugi dio ovog istraživanja je analiza podataka, izgradnja modela predviđanja algoritmom slučajnih šuma te interpretacija ishoda. Tumačenje rezultata našeg modela pomoću četiri različite tehnike glavni je cilj trećeg dijela ovog rada.

U završnom dijelu rada predstavljeni je početni eksperiment u kojemu je primijenjena ista metodologija na specifičnom slučaju predviđanja zrelosti breskvi mjerenjem impedancije na pojedinim voćkama, a sa naglaskom na interpretaciju rezultata korištenog "black box" algoritma [23].

III. Pregled istraživanja

A. Dubinska analiza podataka s područja edukacije (engl. Educational Data Mining)

Stvaranje preciznog modela koji može predviđati ponašanje učenika ili njegove konačne ocjene na temelju njegove aktivnosti vrlo je privlačno bilo kojoj obrazovnoj ustanovi.

Kako bi klasificirali studente koji su pali ispit, Yukselturk i sur. (2014.) koristili su četiri algoritma dubinske analize podataka; k-Najbližih susjeda, stablo odluke, naivni Bayes i neuronske mreže. Tri su se varijable u njihovim konačnim rezultatima pokazale kao najvažniji faktori u predviđanju pada ispita; *samoefikasnost mrežnih tehnologija*, *spremnost internetskog učenja* i *prijašnje online iskustvo* (Yukselturk, Ozekes, and Türel 2014).

U drugom provedenom istraživanju, autori su ispitali aktivnosti učenika prema spolu i po vremenu prijave, koristeći LMS Moodle logove aktivnosti. Otkrili su da postoji značajna povezanost; studentice su bile aktivnije i uspješnije u tečaju od muških studenata, a općenito studenti su bili najaktivniji u ispitnim tjednima, posebno na dan prije testova (Kadoic and Oreski 2018).

Mishra i sur., izgradili su model predviđanja uspješnosti na temelju socijalne integracije studenata, akademske integracije i različitih emocionalnih vještina. Ključni utjecaji na rezultate semestra bili su rezultati prethodnog semestra, nakon čega su slijedili dobri akademski rezultati. Od svih emocionalnih svojstava na rezultate semestra utjecalo je samo vodstvo (leadership) i nagon (drive) studenata (Mishra, Kumar, and Gupta 2014).

Koristeći metodologiju dubinske analize podataka koja se temelji na CRISP-DM metodologiji, Chalaris i sur. (2014.) utvrdili su da se na teorijskim tečajevima razumijevanje studenta uglavnom odnosi na instruktora i učinkovitost predavanja, dok se u tečajevima laboratorijske prakse laboratorijske mogućnosti najviše podudaraju s postizanjem ciljeva učenja (Chalaris et al. 2014).

Predviđanje neuspjeha učenika ili otkrivanje koji faktori najviše utječu na pad studenata u MOOC-u (Gupta and Sabitha 2019) mogu pomoći nastavnicima da redizajniraju značajke MOOC-a (Xing 2019), personaliziraju nastavne procese (Zhang et al. 2019), povećaju uspješnost učenika (Ajibade, Bahiah Binti Ahmad, and Mariyam Shamsuddin 2019) i u konačnici spriječe studente da napuste tečaj.

Naravno, istraživanje podataka o studentima mora se provoditi na etički način, poštujući njihovu privatnost.

B. GDPR i rudarenje podataka

Opća uredba o zaštiti podataka iz 2018., poznata kao GDPR, najvažnija je promjena u propisima o privatnosti podataka u 21. stoljeću. Značajno utječe na mnoge aspekte prikupljanja i obrade podataka građana EU-a i utječe ne samo na tvrtke u EU-u, već i na multinacionalne kompanije koje posluju u EU. Modele strojnog učenja pogoni velika količina osobnih podataka. To znači da privatnost pojedinca moramo poštivati na etički način kako bismo izbjegli rizike vezane uz privatnost (Ashford 2019).

"Pravo na objašnjenje" je još jedan značajan učinak GDPR-a na strojno učenje. Prema Gregoryju Piatetskyu GDPR zapravo ne zahtijeva objašnjenje algoritama strojnog učenja, već on razlikuje *globalno objašnjenje* i *lokalno objašnjenje*. Globalno objašnjenje uglavnom je usredotočeno na funkcioniranje algoritma strojnog učenja. S druge strane, lokalno objašnjenje bavi se pitanjem faktora koji su pridonijeli donošenju određene odluke (Piatetsky-Shapiro 2018).

Kada govorimo o "smislenom objašnjenju" o logici algoritma, moramo gledati iz perspektive subjekta podataka. Ovdje otkrivanje punog koda algoritma i detaljni tehnički opisi procesa strojnog učenja vjerojatno ne bi bili od velike pomoći. Dok bi s druge strane, jednostavni, ne-tehnički opis postupka vjerojatno bio smisleniji. U istom članku autori postavljaju pitanje jednog od zahtjeva GDPR-a da na automatizirane odluke čovjek treba imati pravo na žalbu. Treba li taj zahtjev izvrnuti? Treba li stroj imati pravo žalbe na odluke čovjeka(Kuner et al. 2017)?

IV. Opis skupa podataka

Skup podataka upotrijebljen u ovom istraživanju sadrži 408 zapisa prikupljenih od 5 generacija aktivnosti studenata koji su pohađali predmet "Programiranje 2". Skup podataka sadrži 6 varijabli: *ID*, *lectures*, *quizzes*, *labs*, *videos* i *score*. *ID* varijabla predstavlja učenika; iako je skup podataka anoniman, ova je varijabla uklonjena. *Lectures*, *quizzes* i *labs* varijable predstavljaju ukupan broj bodova koje su studenti dobili na predavanjima, kvizovima i vježbama. Varijabla *videos* predstavlja broj pregleda videa sa predavanjima, a *score* predstavlja ocjenu studenta na završnom ispitu. Podaci za ovo istraživanje prikupljeni su kako je opisano u prethodnom istraživanju, gdje su korištene interpretabilne neuronske mreže u predviđanju neuspjeha studenata (Matetic 2019). Uzorak podataka upotrijebljen u ovom istraživanju prikazan je u Tablici 1.

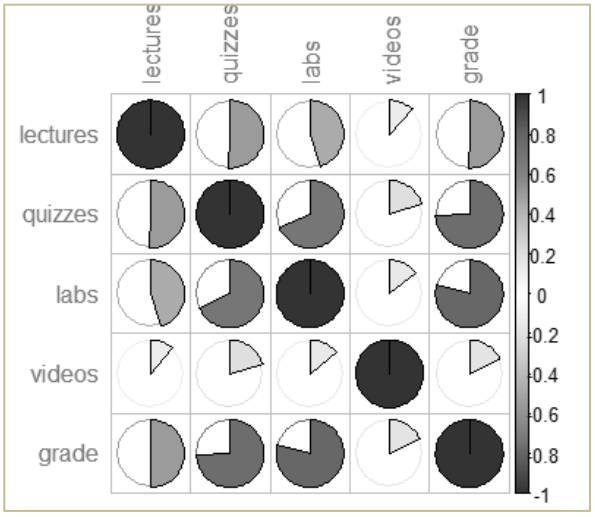
TABLICA I. UZORAK SKUPA PODATAKA

lectures	quizzes	labs	videos	grade
0	19,33	32	15	D
5	22	27	7	D
5	15	7	10	F
5	27,66	27,5	13	C
3	28,66	0	50	F

V. ISTRAŽIVAČKA ANALIZA PODATAKA

Prvi korak, koji prethodi izgradnji modela predviđanja, jest istraživanje podataka.

Pokušavamo predvidjeti ocjenu (*scores*), stoga moramo obratiti pažnju na varijable *labs* i *quizzes* koje imaju najjaču korelaciju s varijablom *scores*. Kako sugerira podatkovna *toplinska mapa* (Slika 1.), *labs* i quizzes imaju najjaču korelaciju, a *videos* i *lectures* najslabiju.

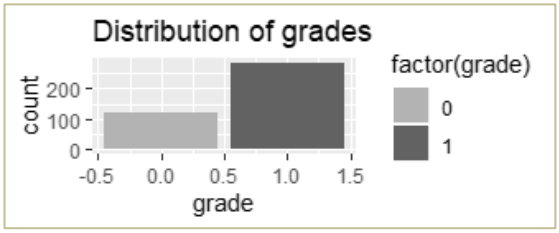


Slika 1. Korelacije između varijabli

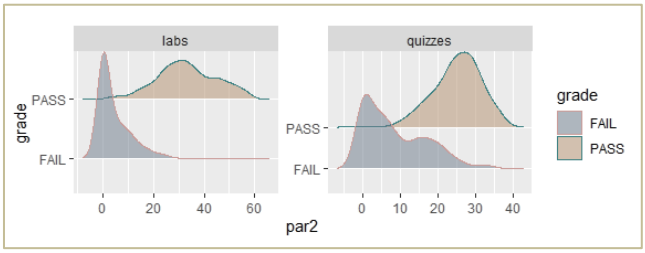
Grafikon na Slici 2. pokazuje da broj ocjena PASS ima znatno više od ocjene FAIL. To znači da se za bolje rezultate podaci moraju normalizirati.

Analizom grafikona na Slici 3, iz raspodjele ocjena učenika (FAIL ili PASS); možemo vidjeti da slabiji rezultati u laboratorijima i kvizovima uglavnom rezultiraju neuspjehom, što nam daje desno iskrivljenu normalnu distribuciju. To je nešto što smo očekivali.

Zanimljiva činjenica koju možemo primijetiti na grafikonu kviza jest da je distribucija FAIL-a blago bimodalna, što nam pokazuje da određeni broj studenata s relativno dobrim rezultatima na kvizovima i dalje padaju ispit.



Slika 2. Distribucija ocjena (0-FAIL, 1-PASS)



Slika 3. Raspodjela ocjena studenata (FAIL i PASS) po *labs* i *quizzes* varijablama; x os prikazuje bodove aktivnosti učenika

VI. IZGRADNJA MODELA PREDVIĐANJA KORIŠTENJEM RANDOM FOREST ALGORITMA

Algoritam slučajnih šuma konstruira svako stablo koristeći drugačiji uzorak podataka i mijenja način konstrukcije klasifikacijskog ili regresijskog stabla (Liaw and Wiener 2003).

Dok se u standardnim stablima, svaki čvor grana koristeći najbolji razdjelnik među svim varijablama, algoritam slučajnih šuma grana svaki čvor koristeći najbolji među podskupom prediktora koji su nasumično odabrani na tom čvoru. Ova metoda funkcionira vrlo dobro u usporedbi s mnogim drugim metodama, uključujući diskriminantnu analizu, metodu potpornih vektora i neuronske mreže, dok je istovremeno poprilično otporna na prenaučenost (engl. *overfitting)*. Vrlo je korisna u smislu da ima samo dva parametra - broj varijabli u slučajnom podskupu na svakom čvoru i broj stabala u šumi, a obično nije vrlo osjetljiva na njihove vrijednosti (Breiman 2001).

Prvi korak u našem procesu bio je dijeljenje naših podataka u dva skupa: podaci za treniranje (80%) i testni podaci (20%). Koristili smo trostruku unakrsnu validaciju ponovljenu 5 puta, a zatim smo izradili model slučajnih šuma (rf_model) sa centriranim i skaliranim podacima. Nakon izrade modela testiran je na testnim podacima, a točnost modela bila je 96,3%.

Dakle, koristeći algoritam slučajnih šuma izgradili smo model koji poprilično precizno predviđa uspjeh studenata, ali nemamo pojma kako on to radi.

Random forest algoritam je tzv. "black box" algoritam. Takvi modeli nam daju malo informacija o postupcima donošenja odluka, pa nam je potreban dodatni napor da bismo to objasnili (Grigg 2019).

VII. INTERPRETACIJA RANDOM FOREST MODELA

Algoritmi koji korisniku kriju svoju unutarnju logiku, takozvani "black box" algoritmi, daju nam malo informacija o njihovim postupcima odlučivanja. Ovaj nedostatak objašnjenja predstavlja praktično i etičko pitanje. Postoje mnogi pristupi kojima se želi prevladati ta slabost po cijenu smanjenja točnosti, a u korist interpretabilnosti (Guidotti et al. 2018).

S druge strane, modeli koje je lako interpretirati (whitebox) kao što su linearna regresija i stabla odlučivanja obično su netočni, jer često ne uspijevaju zabilježiti složene odnose unutar skupa podataka. U ovom radu korišteno je nekoliko metoda za tumačenje rezultata našeg modela slučajnih šuma.

1) Određivanje važnosti varijabli (eng. variable importance)

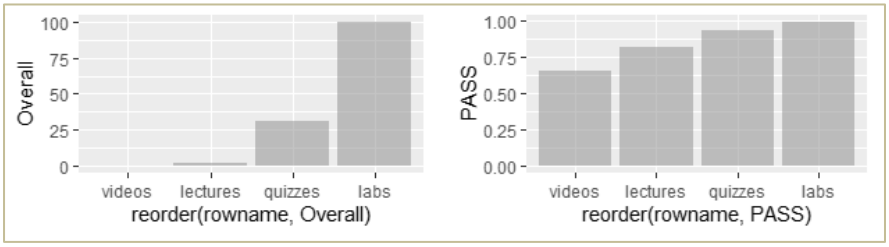
Kod izrade modela slučajnih šuma, normalno je pitati se koje varijable imaju najviše prediktivne moći. Varijable visoke važnosti ključne su za izradu modela i njihove vrijednosti značajno utječu na vrijednosti ishoda. S druge strane, varijable male važnosti mogu se izostaviti iz modela te ga pojednostaviti i ubrzati (Hoare 2019).

Kao što možemo vidjeti na grafikonima važnosti varijabli (Slika 4.), varijabla *labs* najvažnija je varijabla u procesu donošenja odluka. Prediktor v*ideos* relativno je važan za vrijednost klase PASS, ali u cjelini je nebitan.

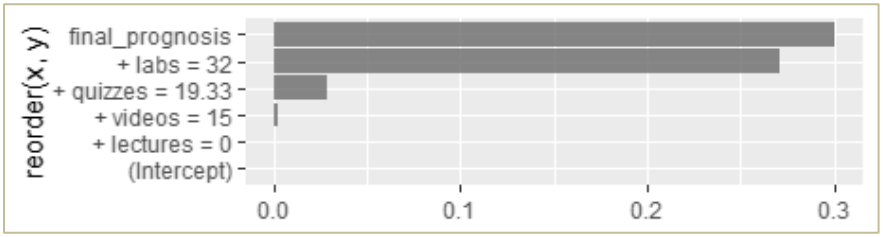
2) Break Down model

Break Down je model agnostičke metode za dekompoziciju predviđanja "black box" algoritama poput slučajnih šuma, Xgboost, metode potpornih vektora ili neuronske mreže. Kao rezultat dobivamo dekompoziciju predviđanja modela koja se može pripisati određenim varijablama. Break Down grafikon vizualno prikazuju njihove doprinose (Slika 5.).

Koristeći R kôd s breakDown paketom, otkrili smo varijable koje su najviše pridonijele našem konačnom predviđanju. Ova metoda daje nam istu varijablu *labs* kao najvrjedniji prediktor, a što odgovara rezultatu alata za određivanje važnosti varijabli.



Slika 4. Ukupna važnost varijabli i važnost varijabli za PASS



Slika 5. Break down grafikon vizualizira doprinose varijabli rezultatima modela

3) Tree surrogate

Tree surrogate metoda koristi jednostavniji model stabla odlučivanja kako bi objasnila kompliciraniji model. Ova metoda kreira stablo odlučivanja na ulaznim podatcima modela i na njegovim predviđanjima, dajući tako sloj preglednosti i objašnjivosti modela.

Vrijednost R-kvadrata (objašnjena varijanca) je mjera koliko dobro model odgovara podatcima ili koliko ih dobro opisuje. Naš surogat model ima vrijednost R-kvadrata od 0,836, što znači da prilično dobro opisuje ponašanje crne kutije, ali ne i savršeno. Kao što možemo vidjeti na Tree surrogate grafici (Slika 6.) varijabla *labs* je opet najvažniji prediktor. Na desnoj strani grafikona se vidi da su varijable *labs* i *quizzes* najviše doprinijele predviđanju vrijednosti klase PASS, dok su prediktori *labs* i *lectures* (lijeva strana grafikona) najvažniji za predviđanje vrijednosti klase FAIL.

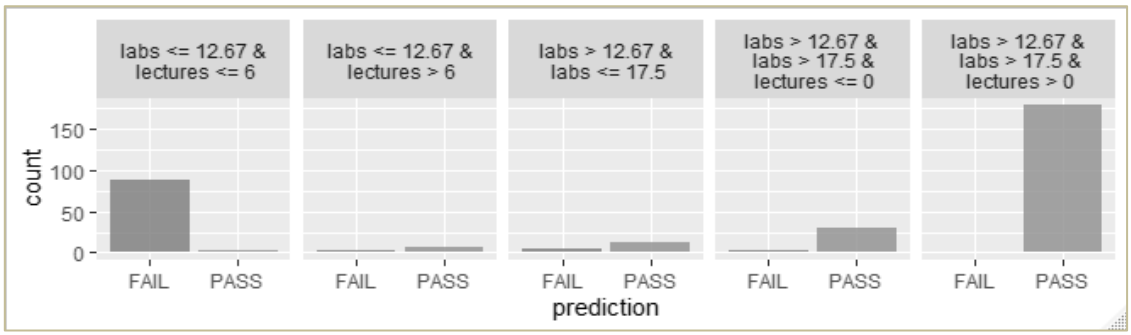
Rezultati se dobiveni u obliku stabla odlučivanja, što je lako protumačiti za razliku od Random forest algoritma koji nema transparentnost.

4) Local Interpretable Model-agnostic Explanations (LIME)

LIME je interpretabilna tehnika koja lokalno oko predviđanja uči interpretativni model koji na objašnjiv i vjeran način opisuje predviđanja bilo kojeg klasifikatora (Guidotti et al. 2018).

LIME objašnjava predviđanja "black box" klasifikatora na način da je za svako dano predviđanje i bilo koji klasifikator u mogućnosti odrediti mali skup značajki u izvornim podacima koji su doveli do rezultata predviđanja. Stvara se model lokalno vjernog agnostičkog skupa objašnjenja koji nam pomaže da razumijemo kako prvobitni model donosi svoju odluku. Stvaranjem reprezentativnog skupa uzoraka LIME korisnicima pruža globalni prikaz granice modela odluke.

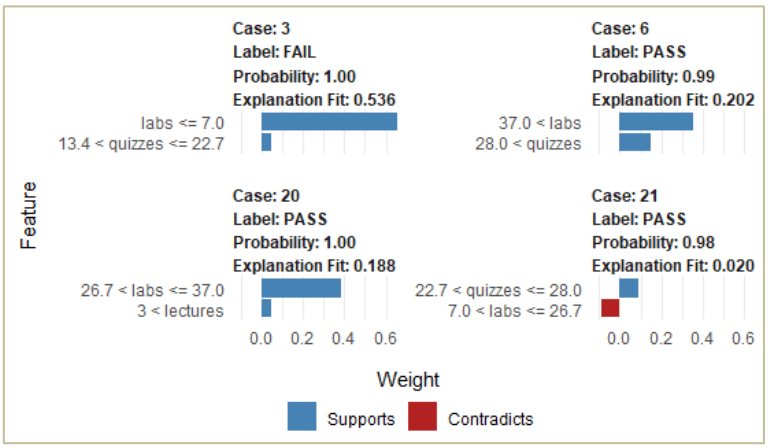
R će nam dati izlazni kod (Slika 7.) s ogromnim brojem pojedinačnih ishoda koje individualno predviđaju prediktori u njihovoj okolini. Ta objašnjenja mogu se vizualizirati, ali bi završili sa ogromnim popisom slučajeva i njihovih grafikona. Slika 8. prikazuje nam samo mali uzorak vizualiziranih objašnjenja (slučajevi od 3 do 10 od 172).



Slika 6. Grafikon Tree surrogate metode

model_type	case	label	label_prob	model_r2	model_intercept	model_prediction	<chr>
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>		
1 classific~	3	FAIL	1	0.539	0.193	0.812	
2 classific~	3	FAIL	1	0.539	0.193	0.812	
3 classific~	6	PASS	0.992	0.206	0.585	1.08	
4 classific~	6	PASS	0.992	0.206	0.585	1.08	
5 classific~	10	PASS	1	0.0234	0.675	0.673	

Slika 7. R izlazni kod - Uzorak pojedinačnih slučajeva sa odgovarajućim prediktorima i njihovim težinama (LIME)



Slika 8. LIME uzorak grafičkih objašnjenja

VIII. SLUČAJ PREDIKCIJE ZRELOSTI BRESKVI I INTERPRETACIJA REZULTATA

A. *Opis problema*

U sklopu projekta "Dubinska analiza tokova podataka za pametno upravljanje hladnim lancem (SmaCC)" Sveučilišta u Rijeci, analiziraju se podaci mjereni na stotinjak plodova breskve sa ciljem predikcije zrelosti voća na osnovi mjerenih parametara (Wang et al. 2017). Dobiveno je desetak parametara svake pojedine voćke: *visina*, *širina*, *radijus*, *volumen*, *masa*, *gustoća*, *tvrdoća*, *težinski postotak šećera* (engl. soluble solids concentration, *SSC*), *količina kiseline u plodu* (titratable acidity, *TA*) te dva podatka *Zs* i *theta* koji predstavljaju mjeru impedancije (strujnog otpora) svake voćke.

Namjera je da se pronađe model, koji će iz mjerenih vrijednosti pokušati što preciznije predvidjeti zrelost voća, koju u našim slučaju predstavlja parametar *tvrdoća*, koji je dobiven kao prosječna vrijednost od četiri mjerenja tvrdoće svake pojedine breskve.

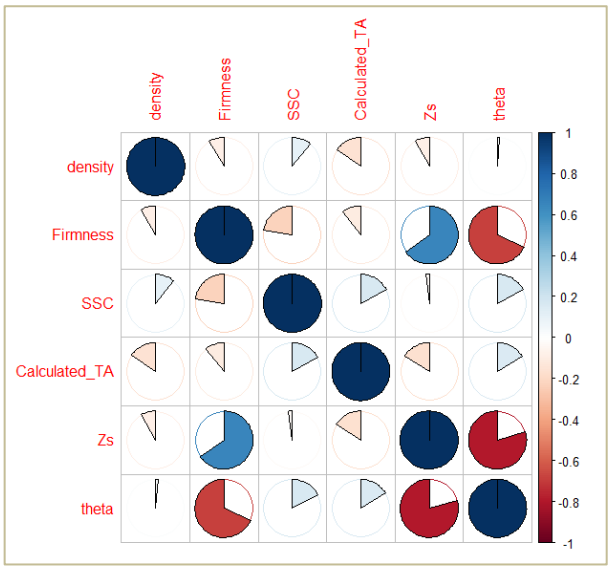
Implementacija funkcionalnog modela bi bila od velike koristi u distribuciji voća jer bi takav postupak predstavljao neinvazivnu metodu određivanja zrelosti, za razliku od klasičnog mjerenja tvrdoće koji svojim postupkom uništava breskvu. Postupak bi se stoga mogao izvoditi veliki broj puta na istim voćkama, što između ostalog, predstavlja i veliku uštedu.

B. *Analiza podataka*

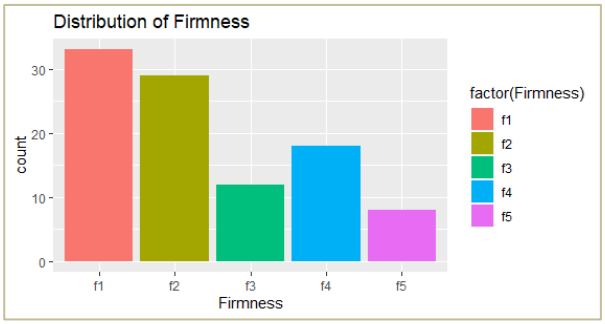
Pretprocesiranje podataka uključivalo je preporuke eksperta, te su iz mjerenog skupa podataka odstranjene varijable: *visina*, *širina*, *radijus*, *volumen*, *masa* pošto se iz tih podataka izračunava *gustoća*.

Kreiran je grafikon korelacije preostalih šest varijabli (Slika 9.) iz čega se odmah može uočiti visoka korelacija između varijable *tvrdoća* i varijabli impedancije *Zs* i *theta*. Međusobnu korelaciju varijabli *Zs* i *theta* nećemo uzimati u obzir, pošto su to dvije značajke impedancije i međusobno jesu u vezi. To što *Zs* i *theta* koreliraju s *tvrdoćom* je dobro, jer to i pokušavamo dokazati.

Tvrdoća breskve ima vrijednosti između 0,33 i 7,63, pa smo napravili diskretizaciju na način da smo tvrdoću od 0 do 1 označili sa f1, od 1 do 2 sa f2, od 2 do 3 sa f3, od 3 do 4 sa f4 a tvrdoću iznad 4 sa f5. Na taj način dobili smo distribuciju *tvrdoće* kao na Slici 10.



Slika 9. Korelacije između varijabli



Slika 10. Distribucija vrijednosti varijable *tvrdoća*

C. *Model predikcije slučajne šume*

Nakon što je skup podijeljen na dva dijela; skup za treniranje 80% te skup za testiranje 20%, kao i u prethodnom primjeru izgrađen je model slučajne šume kad treniranim podatcima.

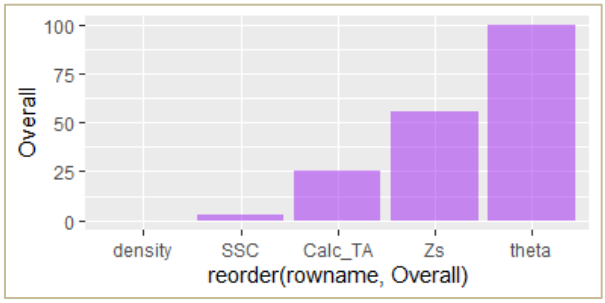
Nakon testiranja na drugom skupu dobili smo točnost predikcije od 47% i slijedeću tablicu konfuzije (Slika 11.). To je mala točnost predikcije, ali nije se ni moglo očekivati bolji rezultat s obzirom na veličinu skupa od svega stotinjak mjerenja.

Prediction	f1	f2	f3	f4	f5
f1	4	3	1	0	0
f2	2	2	1	0	0
f3	0	0	0	1	0
f4	0	0	0	2	1
f5	0	0	0	0	0

Slika 11. Tablica konfuzije

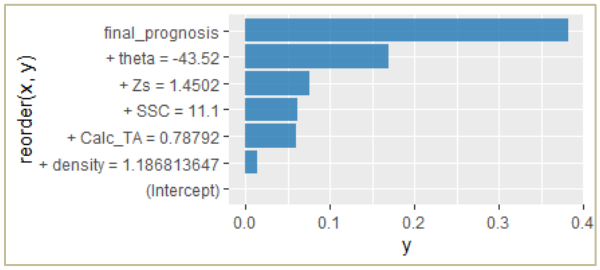
D. *Interpretacija modela*

Prva metoda tumačenja "black box" modela, Određivanje važnosti varijable (engl. variable importance), nam je dala očekivani rezultat te je varijablu *theta* pokazala kao najvažniji input u predikciji (Slika 12.).



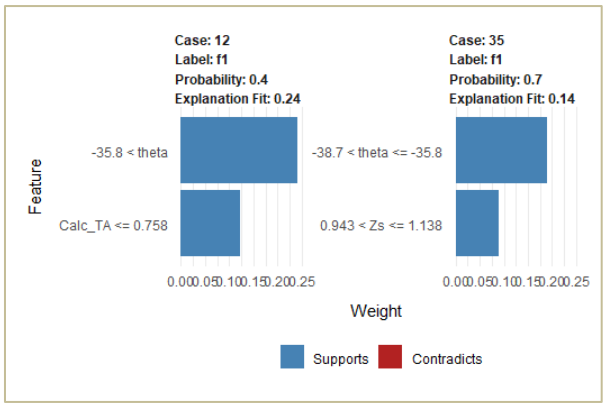
Slika 12. Važnosti varijabli prediktivnog modela

Metodom Break Down dobivamo vrlo sličan rezultat kao i prethodnom metodom, a što je vidljivo na Slici 13.



Slika 13. Break Down grafikon

U LIME metodi smo uzeli samo dva slučaja i kod njih se vidi da čak i lokalno theta ima velik utjecaj na predikciju (Slika 14.)



Slika 14. Slučajevi u LIME interpretativnom modelu

IX. ZAKLJUČAK

Ako nam treba preciznost u predviđanjima, obično smo prisiljeni koristiti modele strojnog učenja koji su uglavnom "black box" modeli. Drugim riječima, ne možemo razumjeti njihove procese učenja ili shvatiti logiku koja stoji iza njegovih zaključaka. Ali postoje alati koji na razumljiv način objašnjavaju granicu našeg modela i u tu svrhu smo koristili nekoliko postupaka u ovom radu.

Ako planiramo poduzeti radnje na temelju predviđanja ili kad odabiremo hoćemo li primijeniti novi model ili ne, bitno je razumjeti razloge koji stoje iza predviđanja, a to je vrlo važno u procjeni povjerenja. Shvaćajući model, možemo transformirati nepouzdan model ili predviđanje u pouzdan.

Kako bismo stvorili povjerenje u naš model, moramo objasniti model ne samo stručnjacima za strojno učenje, već i stručnjacima domena koji zahtijevaju razumljivo ljudsko objašnjenje.

U ovom radu smo koristili algoritam slučajnih šuma kako bismo izgradili model koji može predvidjeti neuspjeh učenika s točnošću od 96,3% što je sasvim dobro, ali ne znajući gotovo ništa o tome koji su ulazi doprinijeli tom rezultatu. Pomoću postupaka za interpretaciju modela otkrili smo dvije najvažnije varijable; *labs* i *quizzes*. Varijabla *labs* je najjači prediktor u svim našim modelima tumačenja i to razumijevanje nam daje priliku da reagiramo u obrazovnom procesu i poboljšamo ga, što je i bio naš početni cilj. Mogli smo koristiti bilo koju od tehnika za tumačenje predviđanja modela, ali postizanje istih rezultata s nekoliko njih daje nam povjerenje u naš model.

Smjernice za budući rad uključuju istraživanje interpretabilnosti modela u domeni upravljanja hladnim lancem u suradnji s ekspertima iz područja, čime se planiraju osigurati uvjeti za pribavljanje dovoljne količine podataka i smjernica za otkrivanje novog i korisnog ekspertnog znanja. Jedan od ciljeva istraživanja je ekspertima primjenom IML metoda predstaviti otkriveno ekspertno znanje na razumljiv i prihvatljiv način.

LITERATURA

[1] Ajibade, Samuel-Soma M, Nor Bahiah Binti Ahmad, and Siti Mariyam Shamsuddin. 2019. "Educational Data Mining: Enhancement of Student Performance Model Using Ensemble Methods." *IOP Conference Series: Materials Science and Engineering* 551: 012061. https://doi.org/10.1088/1757-899x/551/1/012061.

[2] Ashford, Warwick. 2019. "GDPR a Challenge to AI Black Boxes." ComputerWeekly.Com. 2019. https://www.computerweekly.com/news/252452183/GDPR-a-challenge-to-AI-black-boxes.

[3] Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

[4] Chalaris, Manolis, Stefanos Gritzalis, Manolis Maragoudakis, Cleo Sgouropoulou, and Anastasios Tsolakidis. 2014. "Improving Quality of Educational Processes Providing New Knowledge Using Data Mining Techniques." *Procedia - Social and Behavioral Sciences* 147: 390–97. https://doi.org/10.1016/j.sbspro.2014.07.117.

[5] Grigg, Tom. 2019. "Interpretability and Random Forests." Towards Data Science. 2019. https://towardsdatascience.com/interpretability-and-random-forests-4fe13a79ae34.

[6] Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. "Local Rule-Based Explanations of Black Box Decision Systems," no. May. http://arxiv.org/abs/1805.10820.

[7] Gupta, Shivangi, and A. Sai Sabitha. 2019. "Deciphering the Attributes of Student Retention in Massive Open Online Courses Using Data Mining Techniques." *Education and Information Technologies* 24 (3): 1973–94. https://doi.org/10.1007/s10639-018-9829-9.

[8] Hoare, Jake. 2019. "How Is Variable Importance Calculated for a Random Forest?" DisplayR. 2019. https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest.

[9] Kadoic, Nikola, and Dijana Oreski. 2018. "Analysis of Student Behavior and Success Based on Logs in Moodle." *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, 654–59. https://doi.org/10.23919/MIPRO.2018.8400123.

[10] Kuner, Christopher, Dan Jerker B. Svantesson, Fred H. Cate, Orla Lynskey, and Christopher Millard. 2017. "Machine Learning with Personal Data: Is Data Protection Law Smart Enough to Meet the Challenge?" *International Data Privacy Law* 7 (1): 1–2. https://doi.org/10.1093/idpl/ipx003.

[11] "Learning How to Learn: Powerful Mental Tools to Help You Master Tough Subjects." 2019. 2019. https://www.coursera.org/learn/learning-how-to-learn.

[12] Liaw, Andy, and Mathew Wiener. 2003. "Classification and Regression by RandomForest. R News 2." *R News* 3 (December 2002): 18–22.

[13] Matetic, M. 2019. "Mining Learning Management System Data Using Interpretable Neural Networks," 1282–87. https://doi.org/10.23919/mipro.2019.8757113.

[14] Mishra, Tripti, Dharminder Kumar, and Sangeeta Gupta. 2014. "Mining Students' Data for Prediction Performance." *International Conference on Advanced Computing and Communication Technologies, ACCT*, 255–62. https://doi.org/10.1109/ACCT.2014.105.

[15] Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. "Interpretable Machine Learning: Definitions, Methods, and Applications," 1–11. http://arxiv.org/abs/1901.04592.

[16] Nagrecha, Saurabh, John Z. Dillon, and Nitesh V. Chawla. 2019. "MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable." *26th International World Wide Web Conference 2017, WWW 2017 Companion*, 351–59. https://doi.org/10.1145/3041021.3054162.

[17] Piatetsky-Shapiro, Gregory. 2018. "Will GDPR Make Machine Learning Illegal?" KDnuggets. 2018. https://www.kdnuggets.com/2018/03/gdpr-machine-learning-illegal.html.

[18] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should i Trust You?' Explaining the Predictions of Any Classifier." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-Augu: 1135–44. https://doi.org/10.1145/2939672.2939778.

[19] Wang, Xiang, Maja Matetić, Huijuan Zhou, Xiaoshuan Zhang, and Tomislav Jemrić. 2017. "Postharvest Quality Monitoring and Variance Analysis of Peach and Nectarine Cold Chain with Multi-Sensors Technology." *Applied Sciences (Switzerland)* 7 (2). https://doi.org/10.3390/app7020133.

[20] Xing, Wanli. 2019. "Exploring the Influences of MOOC Design Features on Student Performance and Persistence." *Distance Education* 40 (1): 98–113. https://doi.org/10.1080/01587919.2018.1553560.

[21] Yukselturk, Erman, Serhat Ozekes, and Yalın Kılıç Türel. 2014. "Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program." *European Journal of Open, Distance and E-Learning* 17 (1): 118–33. https://doi.org/10.2478/eurodl-2014-0008.

[22] Zhang, Ming, Jile Zhu, Zhuo Wang, and Yunfan Chen. 2019. "Providing Personalized Learning Guidance in MOOCs by Multi-Source Data Analysis." *World Wide Web* 22 (3): 1189–1219. https://doi.org/10.1007/s11280-018-0559-0.

[23] Ljubobratovic, D., Matetic, M. 2019. "Using LMS Activity Logs to Predict Student Failure with Random Forest Algorithm", accepted for publication, 7th International Conference on the Future of Information Sciences - INFuture 2019.