Pregledni rad
https://doi.org/10.31784/zvr.12.1.1
Datum primitka rada: 16. 5. 2023.
Datum prihvaćanja rada: 16. 2. 2024.

PREDIKTIVNE METODE ANALITIKA UČENJA I RUDARENJA OBRAZOVNIH PODATAKA U VISOKOM OBRAZOVANJU TEMELJENE NA ALGORITMIMA STROJNOG UČENJA

Vanja Čotić Poturić
Dipl. ing. mat, viši predavač, Sveučilište u Rijeci, Tehnički fakultet, Vukovarska 58; Fakultet informatike i digitalnih tehnologija (doktorandica), Radmile Matejčić 2, 51 000 Rijeka, Hrvatska;
e-mail: vcotic@uniri.hr

Sanja Čandrlić
Dr. sc., izvanredna profesorica, Sveučilište u Rijeci, Fakultet informatike i digitalnih tehnologija, Radmile Matejčić 2, 51 000 Rijeka, Hrvatska; *e-mail*: sanjac@inf.uniri.hr

Ivan Dražić
Dr. sc., izvanredni profesor, Sveučilište u Rijeci, Tehnički fakultet, Vukovarska 58, 51 000 Rijeka, Hrvatska;
e-mail: ivan.drazic@riteh.uniri.hr

SAŽETAK

U radu je prezentiran pregled literature u posljednjih pet godina o prediktivnim metodama analitika učenja i rudarenja obrazovnih podataka temeljenim na algoritmima strojnog učenja. Primarni kriterij odabira za analizirane radove bio je identificirati one koji koriste algoritme strojnog učenja za predviđanje ishoda u područjima analitika učenja i rudarenja obrazovnih podataka u kontekstu visokog obrazovanja. Važno je naglasiti da ne postoje univerzalne smjernice ili protokoli za predviđanje ishoda u obrazovanju, uključujući i visoko obrazovanje. Metodologija koja se koristi za takva predviđanja prvenstveno ovisi o ciljanoj varijabli i vrsti korištenih ulaznih podataka. U detaljnu analizu uključeno je 25 radova iz citatnih baza Web of Science CC i Scopus. Pomoću šest istraživačkih pitanja ispitano je što se želi predvidjeti u visokom obrazovanju, koji su ulazni podaci korišteni, koliko je algoritama strojnog učenja korišteno u pojedinom istraživanju i koji su bili najučinkovitiji. Također se ispitalo koje su druge tehnike prediktivnog modeliranja navedene te navodi li se programsko okruženje pomoću kojeg je predikcija izvršena.

Ključne riječi: *analitike učenja, rudarenje obrazovnih podataka, predikcija, strojno učenje*

1. UVOD

U sustavu obrazovanja koriste se sustavi za e-učenje (engl. *Learning Management System*, LMS) prilikom različitih oblika izvođenja nastave, kod klasične nastave koja se nadopunjuje informacijsko-komunikacijskim pomagalima, kod mješovitog učenja te kod učenja na daljinu. Sustavi za e-učenje objedinjuju niz funkcionalnosti koje nastavniku omogućuju izvođenje aktivnosti u online okruženju (dostavljanje materijala za učenje, komunikaciju s učenicima, organiziranje e-aktivnosti, vrednovanje) (Hoić-Božić, Holenko Dlab, 2021). Također pružaju podatke o aktivnostima učenika dostupne u izvješćima koja su ugrađena u sustav za e-učenje. Ta su izvješća najčešće prvenstveno deskriptivne prirode, govore sudionicima što se dogodilo, ali ne i zašto se dogodilo te ne predviđaju ishode niti savjetuju sudionike kako poboljšati svoje rezultate.

Tim se problemima bave dva područja, analitike učenja (engl. *Learning Analytics*, LA) i rudarenje obrazovnih podataka (engl. *Educational Data Mining*, EDM). Navedena područja imaju isti cilj, poboljšati nastavu i proces učenja poboljšanjem procesa evaluacije, razumijevanjem problema edukacije i planiranjem intervencija (Siemens, Baker, 2012). Uobičajene metode analitika učenja i rudarenja obrazovnih podataka dane su u Tablici 1. Predmet istraživanja ovog reda je prva navedena metoda, predikcija, kojoj je cilj predvidjeti vrijednosti ciljane varijable iz poznatih vrijednosti ostalih varijabli.

Tablica 1. Neke od uobičajenih metoda LA i EDM

METODA	PRIMJENA
predikcija	predviđanje uspješnosti studenata
grupiranje	grupiranje studenata na temelju njihovih obrazaca učenja
rudarenje odnosa	identificiranje odnosa u obrascima ponašanja studenata
otkrivanje izdvojenica	detekcija učenika s poteškoćama
analiza društvenih mreža	tumačenje strukture i odnosa u suradničkim aktivnostima
rudarenje teksta	analiza sadržaja foruma, dokumenata, web stranica
faktorizacija nenegativne matrice	procjena vještina učenika

Izvor: prilagođeno prema Romero i Ventura (2013), Calvet Liñán i Juan Pérez(2015)

U nastavku rada detaljnije se opisuju područja rudarenja obrazovnih podataka i analitika učenja, kao i o prediktivno modeliranje i algoritmi strojnog učenja koji se koriste za predviđanja u tim područjima.

1. 1 Rudarenje obrazovnih podataka

Rudarenje podataka (engl. *Data Mining*, DM) je ekstrakcija skrivenih korisnih informacija iz skupa podataka kroz znanstvene analize i metode koje identificiraju trendove podataka i skrivene obrasce unutar njih zadanog skupa podataka, te se kao takvo rudarenje podataka može nazvati otkrivanjem znanja (Azevedo, 2018), (Hussain i sur., 2018).

Rudarenje podataka koje se primjenjuje na obrazovne podatke naziva se rudarenje obrazovnih podataka (Baker, Yacef, 2009). Popularnu definiciju za područje EDM-a predlaže 2018. godine Međunarodno društvo za rudarenje podataka u obrazovanju: „EDM je disciplina u nastajanju koja se bavi razvojem metoda za istraživanje jedinstvenih podataka sve većeg opsega koji dolaze iz obrazovnih okruženja, i koristi te metode za bolje razumijevanje učenika i okruženja u kojima uče".

Ovo se područje bavi razvojem metoda koje otkrivaju znanje iz podataka obrazovnog okruženja (Han i sur., 2011), uključujući rudarenje podataka i strojno učenje, psihometriju i druga područja statistike, vizualizaciju informacija i računsko modeliranje (Romero, Ventura, 2007). Rudarenje obrazovnih podataka koristi se za prepoznavanje izazova učenja, za proučavanje i predviđanje uspješnosti učenika (Asif i sur., 2017; Gasevic i sur., 2014; Kostopoulos i sur., 2018) te za evaluaciju integracije tehnologija u proces učenja (Angeli i sur., 2017).

Tri najčešća problema za koje se koriste prediktivne metode rudarenja obrazovnih podataka su otkrivanje hoće li student položiti ili pasti određeni predmet (Conijn i sur., 2016), predviđanje ocjena određenog ispita ili konačnih ocjena predmeta (Moreno-Marcos i sur., 2018) te identifikacija studenata koji će najvjerojatnije odustati (Márquez-Vera i sur., 2013).

Poboljšanje nastave, procesa učenja i poučavanja, cilj je ne samo područja rudarenja obrazovnih podataka već i područja analitika učenja.

1. 2 Analitike učenja

Analitike učenja uključuju mjerenje, prikupljanje podataka, analizu i izvješćivanje (vizualizaciju) o podacima o učenicima i općenito onima koji uče (Long, Siemens, 2011). Na međunarodnoj konferenciji o analitikama učenja LAK 2011. donesena je sljedeća definicija ovog područja: "Analitike učenja su mjerenje, prikupljanje, analiza i izvješćivanje o podacima o učenicima i njihovim kontekstima, u svrhu razumijevanja i optimiziranja učenja i okruženja u kojima se ono odvija." U ovoj se definiciji javljaju tri ključna elementa: podaci, analiza i akcija (Slika 1). Podaci su skup prikupljenih informacija o učeniku, okruženju učenja, interakcijama učenja i ishodima učenja. Analiza podataka je postupak pomoću kojeg se dobiva uvid na temelju podataka u akcije koje je moguće poduzeti. Krajnji je cilj svakog procesa analitika učenja poduzimanje akcije kako bi se poboljšao proces učenja i poučavanja.

Slika 1. Tri ključna elementa analitika učenja

Izvor: autori

Neki od aktualnih ciljeva analitika učenja su podupiranje učenika u razvoju vještina i strategija cjeloživotnog učenja, pružanje personaliziranih i pravovremenih povratnih informacija studentima o njihovom učenju, podrška razvoju važnih vještina kao što su suradnja, kritičko

mišljenje, komunikacija i kreativnost, podrška kvalitetnom učenju i poučavanju pružanjem empirijskih dokaza o uspjehu pedagoških inovacija.

U oba područja, u rudarenju obrazovnih podataka i analitikama učenja, prediktivno modeliranje postalo je temeljna praksa istraživača, s fokusom uglavnom na predviđanje uspjeha učenika (Brooks, Thompson, 2022).

1. 3 Prediktivno modeliranje

Prediktivno modeliranje je skupina tehnika koje se koriste za donošenje zaključaka o neizvjesnim budućim događajima. U obrazovnoj domeni, može nas zanimati predviđanje mjerenja učenja (npr. akademski uspjeh učenika, ili stjecanje vještina), podučavanja (npr. utjecaj datog stil podučavanja ili određenog nastavnika) ili druga predviđanja korisna organizacijama (npr. predviđanja zadržavanja na fakultetu ili upisa kolegija).

Koraci prediktivnog modeliranja su identifikacija problema, prikupljanje podataka, inženjering značajki, odabir značajki, izgradnja modela te evaluacija modela. Najvažnije značajke svakog koraka navedene su u Tablici 2.

Tablica 2. Koraci prediktivnog modeliranja

IDENTIFIKACIJA PROBLEMA	Potrebno je odabrati problem koji će se ponoviti u budućnosti, u kojem postoje mjerljive karakteristike, jasan ishod, mogućnost intervencije te veliki skup podataka.
PRIKUPLJANJE PODATAKA	Istraživač treba identificirati izlaznu varijablu (npr. konačna ocjena ili razina postignuća), kao i moguće ulazne varijable (npr. spol, ocjena na prethodnoj razini, broj pristupa LMS-u).
INŽENJERING VARIJABLI	Inženjering varijabli je korak pretprocesiranja koji transformira neobrađene podatke u varijable koje se mogu koristiti u prediktivnim modelima.
ODABIR VARIJABLI	Kako bi se izgradio i primijenio prediktivni model, potrebno je odabrati prediktivne (ulazne) varijable koje su u korelaciji s izlaznom varijablom, vrijednošću koju treba predvidjeti. Neki modeli koriste sve dostupne varijable za predviđanje, bilo da su vrlo informativne ili ne, dok drugi primjenjuju neki oblik odabira varijabli kako bi se uklonile neinformativne varijable iz modela.
IZGRADNJA MODELA	Za izgradnju modela koriste se algoritmi strojnog učenja.
EVALUACIJA MODELA	Kako bi se procijenila kvaliteta prediktivnog modela, potreban je testni skup podataka s poznatim oznakama. Predviđanja napravljena od strane modela na testnom skupu mogu se usporediti s poznatim pravim oznakama testnog skupa kako bi se procijenio model.

Izvor: autori

Svrha prediktivnog modeliranja je stvoriti model koji će predvidjeti vrijednosti (ili klasu ako se predviđanje ne bavi numeričkim podacima) novih podataka na temelju opažanja. Temelji se na pretpostavci da se skup poznatih podataka može koristiti za predviđanje vrijednosti

ili klase novih podataka na temelju promatranih varijabli. Za izgradnju modela koriste se algoritmi strojnog učenja.

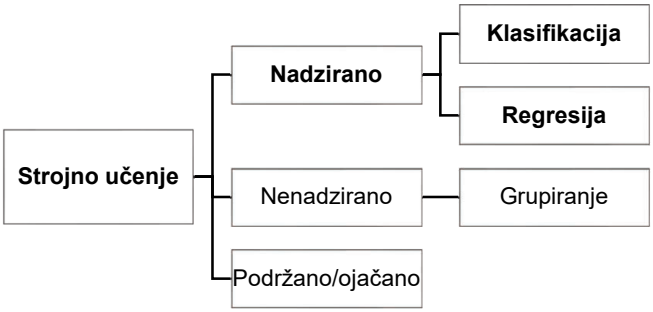
1. 4 Algoritmi strojnog učenja

Strojno učenje (engl. *Machine Learning,* ML) jedno je od danas najaktivnijih i najuzbudljivijih područja računarske znanosti. Ono je grana umjetne inteligencije koja se bave oblikovanjem algoritama koji svoju učinkovitost poboljšavaju na temelju empirijskih podataka. Algoritmi strojnog učenja uče informacije i odnose među njima izravno iz podataka.

Strojno učenje i rudarenje podataka često koriste iste metode i značajno se preklapaju, ali dok se strojno učenje fokusira na predviđanje (temeljeno na poznatim svojstvima naučenim iz trening podataka) rudarenje podataka fokusira se na otkrivanje (prethodno) nepoznatih svojstava u podacima.

Tri osnovna područja strojnog učenja su: nadzirano, nenadzirano i podržano učenje (Slika 2). Nadzirano učenje predviđa vrijednosti izlaznih varijabli na temelju ulaznih podataka. Model se razvija iz trening podataka u kojima su definirane vrijednosti ulaznih i izlaznih varijabli. Model generalizira odnos između ulaznih i izlaznih varijabli te ga koristi za predviđanje drugih skupova podataka u kojima su poznati samo ulazni podaci.

Slika 2. Područja strojnog učenja



Izvor: autori

Dva glavna modela nadziranih metoda su klasifikacijski i regresijski model. Kod klasifikacije, izlazna varijabla je diskretne vrijednosti, a kod regresijskog problema, izlazna varijabla je kontinuirane vrijednosti. Klasifikacijski algoritmi određuju kojoj od predefiniranih kategorija pripadaju ulazni podaci. Zadatak regresijskih algoritama je predviđanje numeričke vrijednosti izlazne varijable nakon zadavanja ulaznih varijabli.

Neki od klasifikacijskih i regresijskih algoritama nadziranog strojnog učenja koji se koriste za izgradnju modela u područjima analitika učenja i rudarenja obrazovnih podataka su: linearna regresija (engl. *Linear Regression,* LR), logistička regresija (engl. *Logistic Regression,* LogR), k-najbližih susjeda (engl. *k-Nearest Neighbors,* kNN), stablo odlučivanja (engl. *Decision Tree,* DT), Naivni Bayesov algoritam (engl. *Naïve Bayes,* NB), stroj potpornih vektora (engl. *Support Vector Machines,* SVM) i umjetna neuronska mreža (engl. *Artificial Neural Network,* ANN).

U nastavku rada drugo poglavlje donosi pregled relevantne literature o istraživanjima u područjima analitika učenja i rudarenju obrazovnih podataka u posljednjih pet godina. Zatim treće poglavlje predstavlja metodologiju odabranu za pregled literature. Četvrto poglavlje prikazuje rezultate s obzirom na definirana istraživačka pitanja te raspravlja o tim rezultatima. Naposljetku, peto poglavlje zaključuje ovaj rad i daje preporuke za buduća istraživanja.

2. PREGLED DOSADAŠNJIH ISTRAŽIVANJA

Ovo poglavlje rada donosi pregled relevantne literature o analitikama učenja i rudarenju obrazovnih podataka u posljednjih pet godina.

Opsežan pregled rudarenja obrazovnih podataka i analitika učenja u visokom obrazovanju (Aldowah i sur., 2019) je najcitiraniji rad u posljednjih pet godina u bazi Web of Science CC za pretraživanje *educational data mining* (Title) AND *learning analytics* (Title). Ovaj pregled obuhvatio je najrelevantnije radove od 2000. do 2017. koji se odnose na četiri glavne računalno podržane analitike: učenja, prediktivne, bihevioralne i vizualizacijske. Na temelju analize 402 rada utvrđeno je da su glavne tehnike rudarenja podataka kao što su klasteriranje, pravilo pridruživanja, vizualno rudarenje podataka, statistika i regresija prikladne za korištenje u područjima LA i EDM. Međutim, ovaj je pregled otkrio da se neke tehnike, kao što su sekvencijalno rudarenje po uzorku, rudarenje teksta, korelacijsko rudarenje, otkrivanje izdvojenih vrijednosti, uzročno rudarenje i procjena gustoće, uobičajeno ne koriste zbog složenosti dobivanja atributa potrebnih za reguliranje ili prilagođavanje individualnim potrebama.

Istraživanje (Romero, Ventura, 2020) dopunjena je i poboljšana verzija prethodnog istraživanja objavljenog 2013. godine. To je najcitiraniji rad u posljednjih pet godina u bazi Scopus za pretraživanje (TITLE (*educational* AND *data* AND *mining*)) AND (TITLE (*learning* AND *analytics*)). Ovaj rad donosi pregled glavnih publikacija, ključnih prekretnica, ciklusa otkrivanja znanja, glavnih obrazovnih okruženja, specifičnih alata, besplatno dostupnih skupova podataka, najčešće korištenih metoda, glavnih ciljeva i budućih trendova u ovom području istraživanja.

Sustavni pregled literature iz 2022. godine (Roslan, Chen, 2022) imao je za cilj identificirati nove trendove i metode korištene za predviđanje akademskog uspjeha studenata u istraživanjima od 2015. do 2021. Autori daju pregled 58 istraživačkih članaka iz baza podataka Lens i Scopus te pokazuju da je istraživački fokus članaka identificiranje čimbenika koji utječu na izvedbu učenika, izvedba algoritama za rudarenje podataka te rudarenje podataka koje se odnosi na sustave e-učenja. Autori također navode da su akademski i demografski podaci primarni faktori koji utječu na uspjeh učenika. Najčešće korišteni pristup je klasifikacija, a klasifikator stablo odlučivanja je najčešće korišteni algoritam.

Autori preglednog rada (Namoun, Alshanqiti, 2021) ističu da je njihovo istraživanje među prvim istraživačkim naporima da se sintetiziraju inteligentni modeli i paradigme primijenjene u obrazovanju za predviđanje postignuća studentskih ishoda učenja, koji predstavljaju zamjenu za studentski učinak. Analizirali su ukupno 62 relevantna rada s fokusom na tri

perspektive, oblike u kojima se predviđaju ishodi učenja, prediktivne analitičke modele razvijene za predviđanje učenja učenika i dominantne čimbenike koji utječu na rezultate učenika. Zaključili su da je postizanje ishoda učenja mjereno uglavnom kao rezultat unutar grupe (tj. rangovi) i rezultat postignuća (tj. ocjene). Regresija i nadzirani modeli strojnog učenja često su korišteni za klasifikaciju uspješnosti učenika. Najbolji prediktori ishoda učenja bili su online aktivnosti učenja učenika, ocjene na polugodištu i akademske emocije učenika.

Glavni cilj istraživanja (Abu Saa i sur., 2019) bio je identificirati najčešće proučavane čimbenike koji utječu na uspjeh učenika i najčešće tehnike rudarenja podataka koje se koriste za identificiranje tih čimbenika. Rezultati su pokazali da su najčešći čimbenici grupirani u četiri glavne kategorije, a to su prethodne ocjene učenika, aktivnosti e-učenja učenika, demografski podaci učenika i socijalni podaci učenika. Osim toga, rezultati su pokazali da su najčešće korištene tehnike rudarenja podataka za predviđanje stablo odlučivanja, naivni Bayesov klasifikator i umjetne neuronske mreže.

Istraživanje (Alyahyan, Düstegör, 2020) pruža nastavnicima jednostavniji pristup tehnikama rudarenja podataka kako bi ostvarili puni potencijal svoje primjene u obrazovanju. Dva najvažnija čimbenika za predviđanje uspjeha studenata, a to su prethodna akademska uspješnost i demografska obilježja studenata, zabilježena su u 69 % istraživačkih radova. Što se tiče tehnika predviđanja, primijenjeni su mnogi klasifikacijski algoritmi za predviđanje uspjeha učenika; algoritmi stabla odlučivanja (44 %), Bayesovi algoritmi (19 %), umjetne neuronske mreže (10 %), algoritmi učenika pravila (9 %), ansambli (7 %), algoritam k-najbližih susjeda (5 %). WEKA je bio najčešće korišteni alat za prediktivno modeliranje. Najčešće korištene evaluacijske mjere su točnost, odziv, preciznost, specifičnost, F-mjera i ROC-krivulja.

Autori rada (Dhankhar i sur., 2021) imali su za cilj provesti sustavni pregled literature o predviđanju uspješnosti učenika pomoću rudarenja obrazovnih podataka i analitika učenja kako bi identificirali tehnike, atribute i mjere koji se koriste. Najčešće korištene tehnike predviđanja bile su stablo odlučivanja, regresija i neuronska mreža. Autori zaključuju da većina promatranih istraživanja ima tendenciju predviđanja uspješnosti prolaza kolegija (prolaz, pad), ocjene na kolegiju, studenta s rizikom od odustajanja/zadržavanja u tradicionalnom/na daljinu/hibridnom okruženju učenja. Najčešće korišteni atributi za predviđanje bili su podaci temeljeni na klikovima, ocjene kolokvija/zadataka/kvizova te demografski podaci.

Rad (Sghir i sur., 2022) pruža sustavni pregled literature u području prediktivnih analitika učenja kako bi se utvrdili trenutni trendovi i napredak u tom području. Rezultati pokazuju da većina promatranih istraživanja koristi prediktivne modele za procjenu uspješnosti učenika i predviđanje onih učenika kojima prijeti neuspjeh ili odustajanje. Što se tiče tehnika koje se koriste za predviđanje, umjetne neuronske mreže, slučajne šume i podizanje nalaze se na prvom, drugom i trećem mjestu, u smislu točnosti predviđanja i učestalosti korištenja u usporedbi s drugim algoritmima. Performanse algoritama obično su se procjenjivale pomoću matrice zabune i mjera dobivenih iz nje.

Vlastitim pregledom literature koji je opisan u sljedećem poglavlju ovog rada žele se pronaći i analizirati relevantni radovi o različitim predikcijama u visokom obrazovanju pomoću

algoritama strojnog učenja kako bi se dao doprinos istraživanjima u područjima analitika učenja i rudarenja obrazovnih podataka.

3. METODOLOGIJA

Svrha ovog pregleda literature je istražiti korištenje prediktivnih metoda u visokom obrazovanju temeljene na algoritmima strojnog učenja koje koriste područja analitika učenja i rudarenja obrazovnih podataka. Svrha će se postići odgovorima na sljedeća istraživačka pitanja:

Q1: Na koji su način prikupljeni podaci za istraživanje, kojeg su tipa prikupljeni podaci te koliko je podataka korišteno u istraživanju?

Q2: Za koju se svrhu koristi predikcija te kojeg je tipa ciljna varijabla?

Q3: Koliko se različitih algoritama strojnog učenja koristi u jednom istraživanju te koji se algoritmi koriste?

Q4: Kojim se mjerama mjeri preciznost/pouzdanost predviđanja?

Q5: Koje dodatne tehnike prediktivnog modeliranja autori navode?

Q6: Navode li autori okruženje (programsku podršku ili programski jezik) pomoću kojeg su izvršili predikciju?

Važno je istražiti koliko je podataka korišteno u istraživanju obzirom na broj studenata i broj promatranih atributa, koji se prediktivni atributi koriste te na koji su način podaci prikupljeni (Q1). Želi se ustvrditi koja je svrha predikcije i kojeg je tipa varijabla koja se predviđa (Q2). O tipu varijable ovisi jesu li korišteni regresijski ili klasifikacijski algoritmi strojnog učenja, koliko je različitih algoritama korišteno u jednom istraživanju te koji se algoritmi koriste (Q3). Na osnovu (Q4) želi se istražiti navodi li se kolika je preciznost predviđanja te način na koji se mjeri. Također je važno istražiti navode li autori dodatne metode prediktivnog modeliranja koje su koristili uz algoritme strojnog učenja poput odabira varijabli, popunjavanja vrijednosti koje nedostaju, metoda preuzorkovanja i sl. (Q5). Postoje različita programska okruženja za strojno učenje te se na osnovu (Q6) želi istražiti navode li autori njihovo korištenje.

Istraživanje je provedeno u prosincu 2022. godine u dvjema bazama, Web of Science CC i Scopus. Pretraživali su se članci iz časopisa i zbornika radova pisani na engleskom jeziku koji imaju otvoreni pristup i objavljeni su između 2018. i 2022. godine. Istraživanje je provedeno prema ključnim pojmovima *„learning analytics"* ili *„educational data mining"* samo u naslovu te *„predict*"* samo u naslovu te *„machine learning"* i *„student*"* u naslovu i/ili sažetku i/ili ključnim riječima.

Upit u bazi Web of Science CC glasio je ((*TI=educational data mining) OR (TI=learning analytics)) AND (TI=predict*) AND (AB=machine learning) AND (AB=student*) AND (PY=(2018 OR2019OR2020OR2021OR2022)) AND (LA=English) NOT (DT=Review article) AND (DT=(Article OR Proceeding Paper)) and Open Access dok je upit u bazi Scopus glasio ((TITLE(learning AND analytics)) OR (TITLE(educational AND data AND mining))) AND (TITLE(predict*)) AND (TITLE-ABS-KEY(student*)) AND (TITLE-ABS-KEY(machine AND learning)) AND (LIMIT-TO(OA ,"all"))*

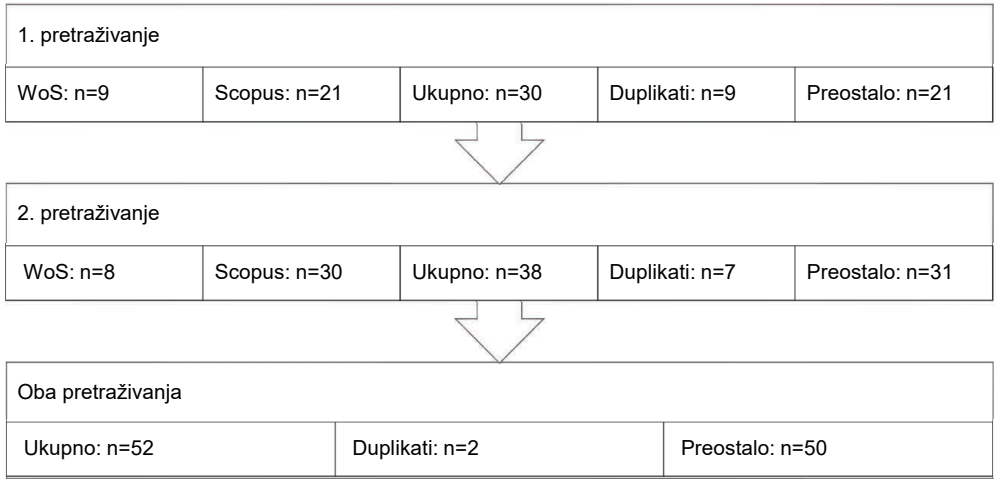
AND (LIMIT-TO(DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO(DOCTYPE, "English") OR LIMIT-TO(DOCTYPE, "2022") OR LIMIT-TO(DOCTYPE,"2021") OR LIMIT-TO(DOCTYPE,"2020") OR LIMIT-TO (DOCTYPE, "2019")).

Kao rezultat pretraživanja za postavljeni upit dobiveno je 9 radova u bazi Web of Science CC te 21 rad u bazi Scopus. U navedenim bazama pronađeno je ukupno 30 radova, od čega je 9 radova bilo u obje baze pa je za analizu preostao 21 rad. Zbog relativno malog broja radova za analizu pretraživanje je prošireno novim upitom prema ključnim pojmovima *„machine learning"* i *„predict*"* i *„student*"* samo u naslovu te *„learning analytics"* ili *„educational data mining"* u naslovu i/ili sažetku i/ili ključnim riječima.

Novi upit u bazi Web of Science CC glasio je (*TI=machine learning predict* student*) NOT (TI=review) AND ((AB=educational data mining) OR (AB=learning analytics)) AND (PY=(2018 OR 2019 OR 2020 OR 2021 OR 2022)) AND (LA=English) NOT (DT=Review article) AND (DT=(Article OR Proceeding Paper)) and Open Access dok je upit u bazi Scopus glasio (TITLE(machine AND learning) AND TITLE(predict*) AND TITLE(student*) AND(TITLE-ABS-KEY(educational AND data AND mining) OR TITLE-ABS-KEY (learning AND analytics)) AND NOT TITLE (review)) AND (LIMIT-TO (OA, "all")) AND (LIMIT-TO(PUBYEAR, 2022) OR LIMIT-TO (PUBYEAR, 2021) OR LIMIT-TO(PUBYEAR, 2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018)) AND (LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English")).*

Kao rezultat pretraživanja za postavljeni upit dobiveno je 8 radova u bazi Web of Science CC te 30 radova u bazi Scopus. U navedenim bazama pronađeno je ukupno 38 radova, od čega je 7 radova bilo u obje baze pa je za analizu preostao 31 rad. Spajanjem rezultata oba pretraživanja dobivena su 52 rada, ali kojih su se 2 rada javila u oba pretraživanja pa je za analizu preostalo 50 radova (Slika 3).

Slika 3. Rezultati pretraživanja baza Web of Science CC i Scopus



Izvor: autori

Nakon čitanja sažetaka eliminirani su radovi koji nisu bitni za ovo istraživanje, različitim kriterijima izbačeno je 25 radova (Tablica 3) pa je naposlijetku za detaljnu analizu preostalo 25 radova koji su navedeni u Tablici 4.

Tablica 3. Broj isključenih radova nakon čitanja sažetaka

KRITERIJ	BR. RADOVA KOJI NE ZADOVOLJAVAJU KRITERIJ	RAZLOG
Rad je napisan na engleskom jeziku.	2	Samo je sažetak na engleskom jeziku.
Rad je posvećen visokom obrazovanju.	10	6 radova se bavi predikcijama u srednjoškolskom obrazovanju te 4 rada predikcijama u MOOC-ovima.
U istraživanju se koriste algoritmi strojnog učenja.	2	Radovi ne koriste algoritme strojnog učenja.
Rad je u otvorenom pristupu.	2	Rad nije u otvorenom pristupu.
Rad nije pregledni rad niti meta-analiza.	6	Isključena su 4 pregledna rada te dvije meta-analize.
Predikcija je vezana za studente.	3	Predikcija je vezana za profesore.
UKUPNO	25	

Izvor: autori

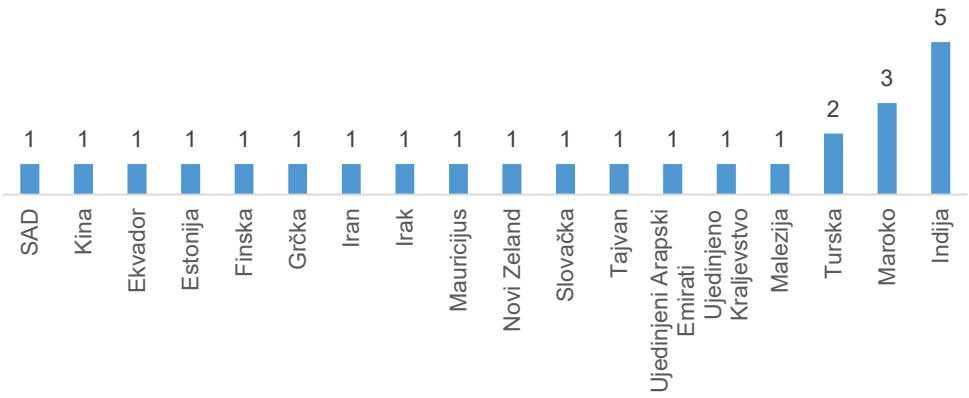
Tablica 4. Radovi uključeni u analizu

Aggarwal i sur. (2019)	Guabassi i sur. (2021)	Lebkiri i sur. (2021)
Badal, Sungkur (2022)	Hashim i sur. (2020)	Niyogisubizo i sur. (2022)
Basu i sur. (2022)	Hooshyar i sur. (2020)	Ouatik F i sur. (2022)
Bayazit i sur. (2022)	Huynh-Cam i sur. (2022)	Ramaswami i sur. (2022)
Buenaño-Fernández i sur. (2019)	Jayapradha i sur. (2019)	Shilbayeh, Abonamah (2021)
Bujang i sur. (2021)	Jeslet i sur. (2021)	Tsiakmaki i sur. (2020)
Dirin, Saballe (2022)	Kabathova, Drlik (2021)	Verma i sur. (2022)
Ghorbani, Ghousi (2020)	Kumar i sur. (2022)	Yağcı (2022)
Gray, Perkins (2019)		

Izvor: autori

Na Slici 4 prikazana je distribucija detaljno analiziranih radova po državama u kojima su provedena istraživanja.

Slika 4. Distribucija analiziranih radova po državama u kojima je provedeno istraživanje



Izvor: autori

4. REZULTATI I RASPRAVA

Ovo poglavlje rada prikazuje raspravu o prikupljenim rezultatima pregleda literature obzirom na definirana istraživačka pitanja.

Q1: Na koji su način prikupljeni podaci za istraživanje, kojeg su tipa prikupljeni podaci te koliko je podataka korišteno u istraživanju?

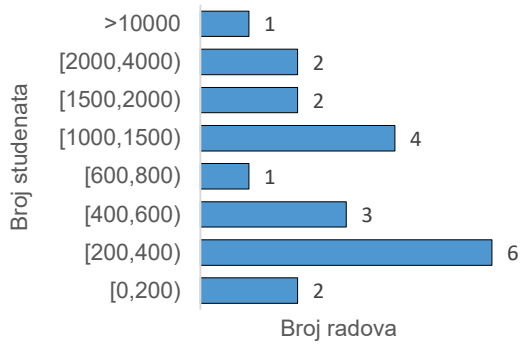
Razlikuju se tri izvora prikupljanja podataka, baza studentskih podataka, sustav za e-učenje i upitnik. U 14 radova korišteni su podaci fakultetske baze, u 3 rada podaci sustava za e-učenje te u 2 rada podaci prikupljeni upitnikom. Dva različita izvora prikupljanja podataka navedena su u 6 radova, od kojih 4 navode fakultetsku bazu podataka i sustav za e-učenje, a 2 fakultetsku bazu podataka i upitnik.

Za predviđanja u visokom obrazovanju koriste se četiri vrste ulaznih podataka, demografski podaci (kao što su spol i dob), podaci prije upisa (kao što su ocjene na prethodnoj razini obrazovanja), podaci nakon upisa na fakultet (kao što su bodovi ili ocjena iz pojedinog kolegija ili aktivnosti) te podaci prikupljeni digitalnim putem (poput vremena provedenog na pojedinoj e-aktivnosti i sudjelovanja u forumima za raspravu). U 8 radova navedeno je korištenje samo jedne vrste ulaznih podataka od kojih 5 radova navode korištenje podataka nakon upisa, 2 rada podatke prikupljene digitalno te 1 rad podatke prije upisa na fakultet. Autori 11 radova navode korištenje dvije vrste ulaznih podataka od kojih se u 6 radova koriste demografski i podaci prije upisa, u 3 rada demografski i podaci nakon upisa na fakultet te u 2 rada podaci prikupljeni digitalnim putem i podaci nakon upisa. U 3 rada navedeno je korištenje tri vrste ulaznih podataka (demografskih, prije i nakon upisa na fakultet) te u 3 rada korištenje svih vrsta ulaznih podataka.

U većini analiziranih radova podaci su dobiveni iz fakultetske baze te su uključivali demografske podatke i podatke o prethodnoj razini studija. Ovi rezultati podupiru istraživanja (Roslan, Chen, 2022) i (Alyahyan, Düstegör, 2020).

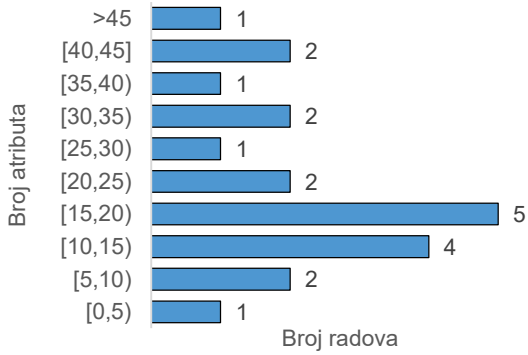
Od 25 analiziranih radova, u 4 rada se ne navodi količina korištenih podataka tj. i broj instanci (studenata) i broj obilježja (atributa). U 3 rada korišteno je više skupova podataka te se za te radove navodi skup s najvećim brojem podataka. Distribucija broja studenata čiji su podaci korišteni za istraživanje prikazana je na Slici 5. Najveći broj studenata je 11 001, a najmanji broj studenata je 69. Distribucija broja obilježja prikazana je na Slici 6. Najmanji broj obilježja je 4, a najveći broj obilježja je 68. Treba istaknuti da je to samo početni broj atributa i da su kasnije uklonjeni neinformativni atributi te su u prediktivnom modelu korištena samo 4 informativna od početnih 68 atributa.

Slika 5. Distribucija radova po broju studenata čiji su podaci korišteni za istraživanje



Izvor: autori

Slika 6. Distribucija analiziranih radova po početnom broju atributa



Izvor: autori

Q2: Za koju se svrhu koristi predikcija te kojeg je tipa ciljna varijabla?

Razlikuju se dvije vrste varijabli, kvalitativne ili kategoričke i kvantitativne ili numeričke. U jednom radu promatraju se obje vrste ciljne varijable, u 2 je rada ciljna varijabla numerička, dok je u preostala 22 rada ciljna varijabla kategorička. Među kategoričkim varijablama razlikuju se

nominalne i ordinalne varijable. Nominalna varijabla koja ima samo dvije neuređene kategorije (prolaz/pad, 0/1, uspjeh/neuspjeh) naziva se dihotomna varijabla i takve se ciljne varijable javljaju u najvećem broju istraživanja, njih 13. Ordinalne varijable s uređenim kategorijama javljaju se kao ciljna varijabla u 7 radova, a u 2 se rada promatraju i dihotomna i ordinalna ciljna varijabla. Samo jedan rad ne navodi tip kategoričke ciljne varijable.

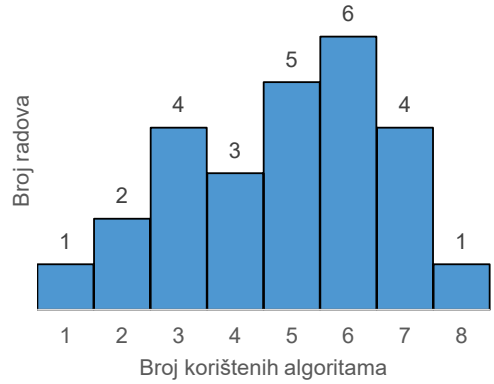
Najveći broj istraživanja, njih 8, želi predvidjeti rizične studente odnosno studente koji su u opasnosti da neće položiti kolegij. Ciljna varijabla u 6 analiziranih radova je konačna ocjena studenata koja je ordinalna varijabla s uređenim kategorijama. Predviđanje prikladnog akademskog programa odnosno studijskog smjera je ciljna varijabla u 3 istraživanja. Hoće li studenti odustati od studiranja ispituje se u 2 istraživanja. Po jedan rad predviđa broj upisanih studenata, hoće li primljeni student upisati fakultet, vjerojatnost da student bude primljen te ponašanje studenta (odgađa li izvršavanje obaveza). Jedno istraživanje ima dvije različite ciljne varijable (konačna ocjena i razina angažmana), a jedno istraživanje tri različite ciljne varijable (konačna ocjena, rizični studenti i odustajanje od studiranja).

Autori (Romero, Ventura, 2020) navode trenutne teme od interesa istraživačke zajednice EDM/LA. Jedan od njih su sustavi ranog upozoravanja ili predviđanje uspješnosti učenika i prepoznavanje rizičnih učenika što je prije moguće kako bi se rano interveniralo i promicao uspjeh učenika. Također, rezultati (Dhankhar i sur., 2021) i (Sghir i sur., 2022) pokazuju da većina postojećih publikacija koristi prediktivne modele za procjenu uspješnosti učenika i predviđanje onih koji su u opasnosti od neuspjeha ili odustajanja. To je u skladu s našim rezultatima - većina istraživanja, njih osam, bavi se predviđanjem rizičnih studenata, nakon čega slijedi učestalost predviđanja konačne ocjene studenata u šest radova.

Q3: Koliko se različitih algoritama strojnog učenja koristi u jednom istraživanju te koji se algoritmi koriste?

U dva se rada koriste regresijski algoritmi, u 22 rada klasifikacijski, a u jednom radu i klasifikacijski i regresijski algoritmi. Klasifikacija je najčešće korištena tehnika za rješavanje prediktivnih problema i taj je rezultat u skladu s pregledima (Aldowah i sur., 2019), (Roslan, Chen, 2022) i (Alyahyan, Düstegör, 2020). Ukupan broj korištenih algoritama je 126, a distribucija broja korištenih algoritama u pojedinom istraživanju prikazana je na Slici 7.

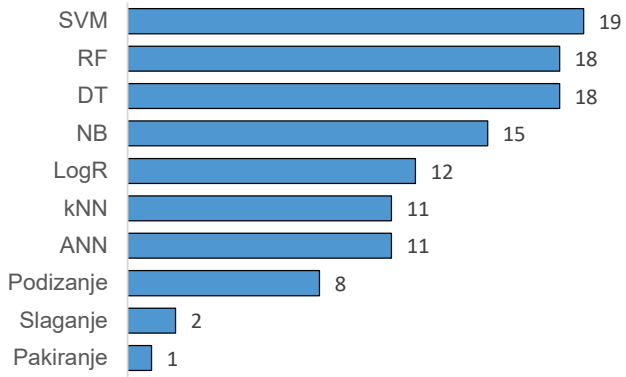
Slika 7. Distribucija analiziranih radova po broju korištenih algoritama strojnog učenja



Izvor: autori

U 23 istraživanja prediktivni zadatak bio je klasifikacijskog tipa te je ukupan broj korištenih algoritama u tim istraživanjima 115. Na Slici 8 prikazana je distribucija korištenih klasifikacijskih algoritama.

Slika 8. Distribucija korištenih klasifikacijskih algoritama u analiziranim radovima



Izvor: autori

Najčešće korišteni algoritmi u analiziranim radovima su stroj potpornih vektora, slučajna šuma, stablo odlučivanja, naivni Bayes i logistička regresija. Korištenje ovih algoritama navodi se i u (Roslan, Chen, 2022), (Abu Saa i sur., 2019) i (Alyahyan, Düstegör, 2020).

Q4: Kojim se mjerama mjeri preciznost/pouzdanost predviđanja?

Ocjenjivanje algoritma strojnog učenja uključuje mjerenje njegove izvedbe na skupu podataka koji se nije koristio, na testnom skupu. Učinkovitost se mjeri pomoću mjere ili skupa mjera koje su specifične za prediktivni zadatak (klasifikacija ili regresija). Svrha evaluacije algoritma je procijeniti njegovu izvedbu na testnim podacima i usporediti njegovu izvedbu s drugim algoritmima ili modelima.

Različiti prediktivni zadaci i podaci mogu zahtijevati različite evaluacijske mjere. Također, u nekim slučajevima koristi se više od jedne mjere procjene kako bi se bolje razumjelo izvođenje modela. U Tablici 5 navedene su evaluacijske mjere korištene u klasifikacijskim algoritmima analiziranih istraživanja, njihov kratak opis te koliko je puta određena mjera korištena.

Tablica 5. Prikaz korištenih evaluacijskih mjera za klasifikaciju

EVALUACIJSKA MJERA	OPIS	BR. KORIŠTENJA
TOČNOST	udio ispravno klasificiranih instanci u skupu podataka	23
PRECIZNOST	udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera	18
ODZIV (OSJETLJIVOST)	udio točno klasificiranih primjera u skupu svih pozitivnih primjera	18
F1	harmonijska sredina preciznosti i osjetljivosti, uravnotežuje obje mjere	14
AUC-ROC	područje ispod ROC krivulje, uspoređuje pravu pozitivnu stopu s lažno pozitivnom stopom	12
TP	broj točnih pozitivnih predviđanja	3
FP	broj netočnih pozitivnih predviđanja	3
KAPPA	pokazuje stupanj slaganja između frekvencija dvaju skupa podataka prikupljenih u dvije različite prilike	2

Izvor: autori

U 15 istraživanja prediktivna varijabla je bila dihotomna te u tim istraživanjima autori kao najučinkovitiji algoritam 4 puta navode LogR, 2 puta RF, 2 puta ansambl Saganje, po jedanput DT, SVM, NB, ansambl Podizanje i Auto-WEKA-u. Autori dvaju radova navode po dva najučinkovitija algoritma, RF i ANN te LogR i SVM. U 11 se radova prilikom navođenja najučinkovitijeg algoritma navodi njegova točnost, u jednom radu AUC-ROC te u 3 rada sve izračunate evaluacijske mjere.

U 9 je istraživanja prediktivna varijabla bila ordinalna te u tim istraživanjima autori kao najučinkovitiji algoritam 4 puta navode RF te jedanput LogR. Autori četiriju radova navode više najučinkovitijih algoritma, 2 puta algoritme RF i ANN te po jedanput ANN i SVM te ANN i SVM i RF i DT. U 8 se radova prilikom navođenja najučinkovitijeg algoritma navodi njegova točnost, a u jednom radu mjera F1.

Q5: Koje dodatne tehnike prediktivnog modeliranja autori navode?

Pretprocesiranje podataka je proces čišćenja, transformacije i organiziranja skupa podataka prije nego što se unese u model strojnog učenja. Ovaj je korak ključan za izvedbu modela,

jer može pomoći u poboljšanju kvalitete podataka i učiniti ih prikladnijima za određeni prediktivni zadatak. Tehnike pretprocesiranja koje se koriste u analiziranim radovima navedene su u Tablici 6.

Optimizacija hiperparametara je proces podešavanja parametara strojnog učenja kako bi se optimizirao njegov rad na određenom zadatku. Uobičajene tehnike za optimizaciju hiperparametara uključuju mrežno pretraživanje, slučajno pretraživanje i Bayesovu optimizaciju. Od 25 analiziranih radova, u njih 5 se navodi optimizacija hiperparametara.

Važnost varijabli u strojnom učenju odnosi se na tehniku određivanja relativne važnosti ulaznih varijabli korištenih u modelu. Na taj se način identificiraju varijable koje imaju najveći utjecaj na predviđanje modela. U analiziranim radovima ova se tehnika spominje 4 puta.

Tablica 6. Prikaz tehnika pretprocesiranja korištenih u analiziranim radovima

TEHNIKA	BR. RADOVA
Odabir informativnih varijabli odnosno uklanjanje neinformativnih varijabli.	19
Transformacija podataka odnosno pretvaranje numeričkih varijabli u kategoričke.	9
Normaliziranje ili standardiziranje podataka.	9
Ponovno uzorkovanje skupa podataka kako bi se smanjio omjer neravnoteže.	6
Ispitivanje koreliranosti varijabli.	6
Zamjena vrijednosti koje nedostaju.	5
Brisanje vrijednosti koje nedostaju.	4
Rukovanje izdvojenicama.	2

Izvor: autori

Q6: Navode li autori okruženje (programsku podršku ili programski jezik) pomoću kojeg su izvršili predikciju?

Autori 8 radova navode korištenje programskog jezika Python, u 7 se radova navodi korištenje softvera WEKA, a u jednom radu korištenje programskog jezika R i softvera WEKA. U 2 se rada navodi korištenje alata Orange, u po jednom radu korištenje alata Hadoop i XLSTAT. Autori jednog rada navode korištenje više alata, IBM SPSS, R, KNIME i Bayesian labs. U jednom se radu koristi vlastita Web aplikacija, dok autori triju radova ne navode korištenje niti jednog okruženja. Ovi rezultati podupiru istraživanje (Alyahyan, Düstegör, 2020) u kojem se navodi da je WEKA najčešće korišten alat za prediktivne modele.

5. ZAKLJUČAK I PREPORUKE ZA BUDUĆA ISTRAŽIVANJA

U radu je opisan pregled literature o prediktivnim metodama u analitikama učenja i rudarenju obrazovnih podataka u visokom obrazovanju na temelju algoritama strojnog učenja. Prilikom odabira radova za analizu, glavni kriterij bio je pronaći radove koji koriste algoritme strojnog učenja za predviđanje u visokom obrazovanju u područjima analitika učenja i rudarenja obrazovnih podataka.

U detaljnu analizu uključeno je 25 radova i postavljeno je šest istraživačkih pitanja. Ispitali smo što se predviđa u visokom obrazovanju, koji su ulazni podaci korišteni, koliko je algoritama strojnog učenja korišteno u pojedinom istraživanju i koji su bili najučinkovitiji. Osim toga, ispitano je koje se druge tehnike prediktivnog modeliranja navode te navode li autori programsko okruženje upotrijebljeno za predviđanje.

Treba naglasiti da ne postoje opća pravila ili postupci za predviđanja u obrazovanju, uključujući visoko obrazovanje. Pristup uglavnom ovisi o ciljnoj varijabli i ulaznim podacima. Krajnji cilj svake analitike učenja je poduzeti pedagoške mjere za poboljšanje procesa učenja i poučavanja. Analizirani radovi ne prikazuju jesu li poduzete pedagoške mjere, niti definiraju teorijski okvir.

Naposljetku, ovo područje treba dodatno istražiti kako bi se pružio širi pedagoško-tehnološki okvir koji bi nastavnicima u sustavu visokog obrazovanja pomogao u izgradnji prediktivnog modela i poboljšanju procesa učenja i poučavanja odgovarajućim pedagoškim intervencijama.

REFERENCE

Abu Saa, A., Al-Emran, M., Shaalan, K., 2019. Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, 24(4), p. 567 – 598.

Aggarwal D., Mittal S., Bali V., 2019. Prediction model for classifying students based on performance using machine learning techniques. *International Journal of Recent Technology and Engineering*, Volume 8, pp. 496-503.

Aldowah H., Al-Samarraie H., Fauzy W. M., 2019. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, pp. 13-49.

Alyahyan, E., Düstegör, D., 2020. Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1).

Angeli, C., Howard, S.K., Ma, J., Yang, J., Kirschner, P.A., 2017. Data mining in educational technology classroom research: can it make a contribution?. *Computers&Education*, Volume 113, pp. 226-242.

Asif, R., Merceron, A., Ali, S.A., Haider, N.G., 2017. Analyzing undergraduate students' performance using educational data mining. *Computers&Education*, Issue 113, pp. 177-194.

Azevedo, A., 2018. Data mining and knowledge discovery in databases.. In: *Encyclopedia of Information Science and Technology.* fourth ed. ed. s.l.:IGI Global, pp. 1907-1918.

Badal Y.T., Sungkur R.K., 2022. Predictive modelling and analytics of students' grades using machine learning algorithms. *Education and Information Technologies*.

Baker, R.,Yacef, K., 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, pp. 3-17.

Basu K., Basu T., Buckmire R., Lal N., 2022. Predictive models of student college commitment. *Education and Information Technologies*, 4(2).

Bayazit, A; Apaydin, N; Gonullu, I, 2022. Predicting At-Risk Students in an Online Flipped Anatomy Course Using Learning Analytics. *EDUCATION SCIENCES*, Volume 12.

Brooks, C., Thompson, C., 2022. Predictive Modelling in Teaching and Learning. In: SoLAR, ed. *Handbook of Learning Analytics.* 2 ed. s.l.:Charles Lang, George Siemens, Alyssa Friend Wise, Dragan Gašević, Agathe Merceron, pp. 29-37.

Buenaño-Fernández D., Gil D., Luján-Mora S., 2019. Application of machine learning in predicting performance for computer engineering students: A case study. *Sustainability (Switzerland)*, 11(10).

Bujang S.D.A., Selamat A., Ibrahim R., Krejcar O., Herrera-Viedma E., Fujita H., Ghani N.A.M., 2021. Application of machine learning in predicting performance for computer engineering students: A case study. *IEEE Access*, Volume 9.

Calvet Liñán, L., Juan Pérez, Á. A., 2015. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. RUSC.. *RUSC. Universities and Knowledge Society Journal*, 3(12), pp. 98-112.

Conijn, R.; Snijders, C.; Kleingeld, A.; Matzat, U., 2016. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Trans. Learn. Technol.*, Volume 10, pp. 17-29.

Dhankhar, A., Solanki, K., Dalal, S. Omdev, 2021. Predicting students performance using educational data mining and learning analytics: A systematic literature review. *Lecture Notes on Data Engineering and Communications Technologies*, Volume 59, p. 127-140.

Dirin A., Saballe C.A., 2022. Machine Learning Models to Predict Students' Study Path Selection. *International Journal of Interactive Mobile Technologies*, Volume 16, pp. 158-183.

Gasevic, D., Rose, C., Siemens, G., Wolff, A., & Zdráhal, Z., 2014. *Learning analytics and machine learning.* s.l., s.n., pp. 287-288.

Ghorbani R., Ghousi R., 2020. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. *IEEE Access*, Volume 8, pp. 67899-67911.

Gray C.C., Perkins D., 2019. Utilizing early engagement and machine learning to predict student outcomes. *Computers and Education*, Volume 131, pp. 22-32.

Guabassi I.E., Bousalem Z., Marah R., Qazdar A., 2021. A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms. *International journal of online and biomedical engineering*, Svezak 17, pp. 135-147.

Han, J., Kamber, M., Pei, J., 2011. *Data mining: Concepts and techniques.* 3 ed. s.l.:Elsevier.

Hashim A.S., Awadh W.A., Hamoud A.K., 2020. *Student Performance Prediction Model based on Supervised Machine Learning Algorithms.* s.l., s.n.

Hoić-Božić N., Holenko Dlab, M., 2021. *Uvod u e-učenje: obrazovni izazovi digitalnog doba.* Rijeka: University of Rijeka, Faculty of Informatics and Digital Technologies.

Hooshyar D., Pedaste M., Yang Y., 2020. Mining educational data to predict students' performance through procrastination behavior. *Entropy*, Volume 22.

Hussain, S., Dahan, N.A., Ba-Alwib, F.M., Ribata, N., 2018. Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(2), pp. 447-459.

Huynh-Cam T.-T., Chen L.-S., Huynh K.-V., 2022. Learning Performance of International Students and Studentswith Disabilities: Early Prediction and Feature Selectionthrough Educational Data Mining. *Big Data and Cognitive Computing*, 6(3).

Jayapradha J., Kumar K.J.J., Deka B., 2019. Educational data classification and prediction using data mining algorithms. *International Journal of Recent Technology and Engineering*, Volume 8, pp. 8674-8678.

Jeslet D.S., Komarasamy D., Hermina J.J., 2021. *Student Result Prediction in Covid-19 Lockdown using Machine Learning Techniques.* s.l., s.n.

Kabathova J., Drlik M., 2021. Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences (Switzerland)*, Volume 11.

Kostopoulos, G., Kotsiantis, S., Pierrakeas, C., Koutsonikos, G., Gravvanis, G.A., 2018. Forecasting students' success in an open university. *Learning Technology*, 13(1).

Kumar T., Sankaran K.S., Ritonga M., Asif S., Sathiya Kumar C., Mohammad S., Sengan S., Asenso E., 2022. Fuzzy Logic and Machine Learning-Enabled Recommendation System to Predict Suitable Academic Program for Students. *Mathematical Problems in Engineering*.

Lebkiri, N; Daoudi, M; Abidli, Z; Elturk, J; Soulaymani, A; Khatori, Y; El Madhi, Y; Benattou, M, 2021. Using Machine Learning for Prediction Students Failure in Morocco: An Application of the CRISP-DM Methodology. *INTERNATIONAL JOURNAL OF EDUCATION AND INFORMATION TECHNOLOGIES*, Volume 15, pp. 344-352.

Long, P., Siemens, G., 2011. Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5), pp. 30-40.

Márquez-Vera, C.; Cano, A.; Romero, C.; Ventura, S., 2013. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Appl. Intell.*, pp. 315-330.

Moreno-Marcos, P.M.; Alario-Hoyos, C.; Muñoz-Merino, P.J.; Kloos, C.D., 2018. Prediction in MOOCs: A review and future research directions. *IEEE Trans. Learn. Technol.*, Svezak 12.

Namoun, A.; Alshanqiti, A., 2021. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Applied Sciences*, Volume 11.

Niyogisubizo J., Liao L., Nziyumva E., Murwanashyaka E., Nshimyumukiza P.C., 2022. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, Volume 3.

Ouatik F., Erritali M., Ouatik F., Jourhmane M., 2022. Predicting Student Success Using Big Data and Machine Learning Algorithms. *International Journal of Emerging Technologies in Learning*, Volume 17, pp. 236-251.

Ramaswami G., Susnjak T., Mathrani A., 2022. On Developing Generic Models for Predicting Student Outcomes in Educational Data Mining. *Big Data and Cognitive Computing*, Volume 6.

Romero, C., Ventura, S., 2007. Educational data mining: A survey from 1995 to 2005. *ScienceDirect*, 33(1), pp. 135-146.

Romero, C., Ventura, S., 2013. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3), pp. 12-27.

Romero, C., Ventura, S., 2020. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3).

Roslan M.H.B., Chen C.J., 2022. Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021). *International Journal of Emerging Technologies in Learning*, 17(5), pp. 147-179.

Sghir, N., Adadi, A., Lahmer, M., 2022. Recent advances in predictive learning analytics: A decade systematic review (2012–2022),. *Education and Information Technologies*.

Shilbayeh S., Abonamah A., 2021. Predicting student enrolments and attrition patterns in higher educational institutions using machine learning. *International Arab Journal of Information Technology*, Volume 18, pp. 562-567.

Siemens, G., Baker, R.S., 2012. *Learning analytics and educational data mining: Towards communication and collaboration.* s.l., s.n., pp. 252-254.

Tsiakmaki M., Kostopoulos G., Kotsiantis S., Ragos O., 2020. Implementing autoML in educational data mining for prediction tasks. *Applied Sciences (Switzerland)*, Volume 10.

Verma S., Yadav R.K., Kholiya K., 2022. A Scalable Machine Learning-based Ensemble Approach to Enhance the Prediction Accuracy for Identifying Students at-Risk. *International Journal of Advanced Computer Science and Applications*, Volume 13, pp. 185-192.

Yağcı M., 2022. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, Volume 9.

Review paper
https://doi.org/10.31784/zvr.12.1.1
Received: 16. 5. 2023.
Accepted: 16. 2. 2025.

THE EMPLOYMENT OF MACHINE LEARNING ALGORITHMS FOR PREDICTION IN LEARNING ANALYTICS AND EDUCATIONAL DATA MINING WITHIN THE CONTEXT OF HIGHER EDUCATION

Vanja Čotić Poturić
Mag. Math., Senior Lecturer, University of Rijeka, Faculty of Engineering, Vukovarska 58; Faculty of Informatics and Digital Technologies (PhD Student), Radmile Matejčić 2, 51000 Rijeka, Croatia, email: vcotic@uniri.hr

Sanja Čandrlić
PhD, Associate Professor, University of Rijeka,Faculty of Informatics and Digital Technologies, Radmile Matejčić 2, 51000 Rijeka, Croatia, email: sanjac@inf.uniri.hr

Ivan Dražić
PhD, Associate Professor, University of Rijeka, Faculty of Engineering, Vukovarska 58, 51000 Rijeka, Croatia, email: ivan.drazic@riteh.uniri.hr

ABSTRACT

This paper presents a review of the literature from the last five years on predictive methods of Learning Analytics and Educational Data Mining based on Machine Learning algorithms. The primary selection criterion for the papers analyzed was to identify those that use Machine Learning algorithms to predict outcomes in the areas of Learning Analytics and Educational Data Mining in the context of higher education. It is important to highlight that there are no universal guidelines or protocols for predicting outcomes in education, including higher education. The methodology used for such predictions depends primarily on the target variable and the type of input data used. Twenty-five papers from the Web of Science CC and Scopus citation databases were included in the detailed analysis. Six research questions were used to examine what is being predicted in higher education, what input data were used, how many Machine Learning algorithms were used in the research, and which were most effective. In addition, the research looked at what other predictive modeling techniques were mentioned and whether the programming environment used for prediction was mentioned.

***Keywords**: Learning Analytics, Educational Data Mining, prediction, Machine Learning*