Educational data mining

Milica Vujić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
vujic.e219.2024@uns.ac.rs

Ivan Partalo
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
partalo.e257.2024@uns.ac.rs

***Apstrakt* — Razumevanje faktora koji utiču na uspeh studenata predstavlja značajan izazov za visoko obrazovanje, jer omogućava bolje planiranje nastavnih procesa i unapređenje kvaliteta studiranja. Tradicionalno, uspeh studenata se posmatra kroz ocene i prisustvo na nastavi, dok se često zanemaruje širi kontekst, poput socio-ekonomskih karakteristika, navika učenja i načina života. Formiranje modela koji može da predvidi uspeh studenata na osnovu različitih faktora od značaja je kako za obrazovne institucije, tako i za same studente. Na ovaj način mogu se identifikovati potencijalni rizici neuspeha, dok studenti mogu steći uvid u aspekte svog ponašanja i okruženja koji značajno doprinose njihovim rezultatima. U ovom radu analizirani su podaci o studentima iz Iraka, Portugala, Pakistana i Južnoafričke Republike, koji obuhvataju akademske rezultate, socio-ekonomske faktore i životne navike studenata, sa ciljem otkrivanja obrazaca i ključnih determinanti akademskog uspeha. Podaci iz različitih izvora objedinjeni su u jedinstveni skup, pri čemu su zadržane zajedničke karakteristike relevantne za sve grupe studenata. Nakon uklanjanja nedostajućih vrednosti i osnovne obrade, formirani dataset je podeljen na trening i test skup, što je omogućilo obuku i evaluaciju modela mašinskog učenja. Za potrebe istraživanja korišćeni su modeli MLP, Random Forest i K-Nearest Neighbors, a njihova tačnost upoređena je na osnovu rezultata testiranja. Dobijeni rezultati omogućili su predikciju budućih rezultata na nivou pojedinačnih studenata. Dodatno, izvršeno je poređenje predikcija za jednog studenta primenom sva tri modela, čime je omogućena procena praktične primenljivosti razvijenih metoda. Ovaj pristup može doprineti boljem razumevanju povezanosti između univerzitetskog okruženja, načina života i akademskog uspeha, kao i razvoju strategija za unapređenje obrazovnih procesa i personalizovanu podršku studentima.**

Ključne reči—rudarenje podataka; ocene; student;, faktori uspeha; klasifikacija

I. UVOD

Razumevanje faktora koji utiču na akademski uspeh studenata omogućava obrazovnim institucijama da unaprede nastavne procese i pruže ciljanu podršku studentima. Predviđanje uspeha studenata na osnovu različitih faktora može pomoći univerzitetima u identifikaciji grupa studenata kojima je potrebna dodatna motivacija i podrška, a studentima da bolje razumeju aspekte svog ponašanja i okruženja koji utiču na akademski učinak.

U ovom radu predstavljeno je rešenje zasnovano na modelima mašinskog učenja za predikciju uspeha studenata, obučeno na podacima prikupljenim iz četiri države – Južnoafričke Republike, Pakistana, Iraka i Portugala. Podaci su objedinjeni u jedinstveni skup, pri čemu su izdvojene zajedničke karakteristike: pol, godina studija, oblast, država, količina učenja, prisustvo nastavi, smeštaj, finansijski status i bliskost sa porodicom. Podaci prisutni samo u nekim dataset-ovima nisu korišćeni kako bi se obezbedila konzistentnost analize.

Izazovi u realizaciji ovog istraživanja uključuju razlike u strukturama podataka, nekompletne informacije i varijacije u načinu merenja uspeha među državama. Faktori koji su prisutni samo u pojedinim dataset-ovima, kao što su nivo stresa, zanimanje roditelja ili konzumacija alkohola, izbačeni su iz analize, jer njihovo uključivanje ne bi omogućilo validno poređenje između studenata iz različitih država.Skup podataka je prethodno pripremljen tako što je standardizovan, spojeni su podaci iz različitih izvora i podeljeni na trening i test skupove, kako bi se omogućila pouzdana obuka i evaluacija modela. Za klasifikaciju studenata u kategorije visoko uspešan, srednje uspešan i slabije uspešan korišćeni su Random Forest, k-NN i višeslojna perceptronska mreža (MLP). Rezultati omogućavaju predikciju uspeha pojedinačnih studenata, analizu ključnih faktora koji utiču na akademski uspeh, kao i uvid u sličnosti i razlike između obrazovnih sistema različitih država.

Detaljniji opis podataka, izazova i metoda analize biće predstavljen u narednim poglavljima. Naredno poglavlje se bavi srodnim istraživanjima na ovu temu. Treće poglavlje obuhvata opis skupova podataka i pripremu podataka za obučavanje i validaciju modela, dok četvrto poglavlje prikazuje metodologiju i rezultate obuke modela. Na kraju sledi diskusija i zaključci istraživanja.

II. SRODNA ISTRAŽIVANJA

U radu [1], istraživači razvijaju modele zasnovane na nadgledanom učenju za predikciju statusa studenata, odnosno da li će student položiti ili pasti završni ispit. Podaci su uključivali atribute poput godine studija, pola, bodova za aktivnosti i rezultate ispita. Korišćeni su algoritmi Decision

Tree, Naïve Bayes, K-Nearest Neighbors (KNN) i Multi-layer Perceptron (MLP), pri čemu su MLP i KNN pokazali visoku tačnost u predikciji konačnih ocena i statusa studenata. Evaluacija je rađena upotrebom TPR, FPR, Precision i Recall, a najbolji rezultati su dostigli 88,6% tačnosti. Na osnovu ovde pokazanih dobrih rezultata, naš rad takođe koristi modele KNN i MLP.

U radu [2], istraživači su koristili podatke studenata osnovnih studija sa Universiti Teknologi MARA u Maleziji, sa ciljem kategorizacije studenata u dve grupe: odličan i nije odličan. Podaci su obuhvatali pol, godine, ukupni prosek i rezultate na ključnim predmetima. Ukupno je uzeto u obzir 631 diplomirani studenat i za svakog je poznato ime, student ID, pol, ukupni prosek (CGPA). GCPA se pokazao kao najvažniji atribut za klasifikaciju. Korišćeni su algoritmi KNN, Decision Tree, Naive Bayes i Logistic Regression. Najveća tačnost postignuta je korišćenjem Naive Bayes algoritma sa 89,26% tačnosti. Takođe smo i mi koristili KNN i tačnost i svojstva pol i prosek.

U radu [3] analizirana je klasifikacija studenata na osnovu različitih životnih i radnih navika, sa ciljem predviđanja uspeha u studijama. Autori su primenom više algoritama, poput OneR, REPTree i DecisionTree, postigli tačnost iznad 76%, pri čemu je uočeno da faktori poput vremena posvećenog učenju i vannastavnih aktivnosti značajno utiču na uspešnost. Posebno je naglašeno da selekcija relevantnih atributa poboljšava performanse modela u odnosu na korišćenje svih dostupnih podataka. Ovi rezultati potvrđuju važnost obrazovnih podataka za unapređenje kvaliteta nastave i zadržavanje studenata u obrazovnom sistemu. Značaj ovog istraživanja ogleda se u tome što daje osnovu za dalje analize, pa je i u našem radu korišćena slična metodologija.

III. Opis skupa podataka

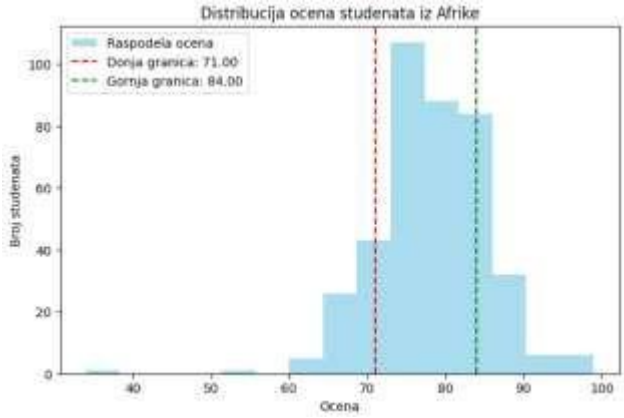
Inicijalno podatke o studentima čine podaci o studentima sa Stellenboch Univerziteta iz Južnoafričke Republike, preuzeti sa [4], podaci o studentima sa Govt Islamia Graduate college Kasur, Punjab, Pakistan, preuzeti sa [5], podaci o studentima iz Porta, Portugal, prikupljeni za rad studije P. Cortez and A. Silva for the study "Using Data Mining to Predict High School Student Performance" preuzeti sa [7], podaci o studentima iz 4 škole iz Iraka preuzeti sa [6]. Iako skupovi podatak sadže dosta faktora studentskog uspeha, problem na koji nailazimo je neusklađenost njihove forme između skupova podataka. Takođe neki faktori koji su nam u jenom skupu značajni, u drugom skupu ih nemamo i nemamo način da ih prikupimo. S toga je obavljena analiza svih skupova i utvrđeno je da su sledeći atributi zajednički:

- pol
- godina studija
- oblast
- država
- količina učenja (sati učenja)
- prisustvo na nastavi
- smeštaj
- finansijski status
- bliskost sa roditeljima.
- ocena (A, B ili C).

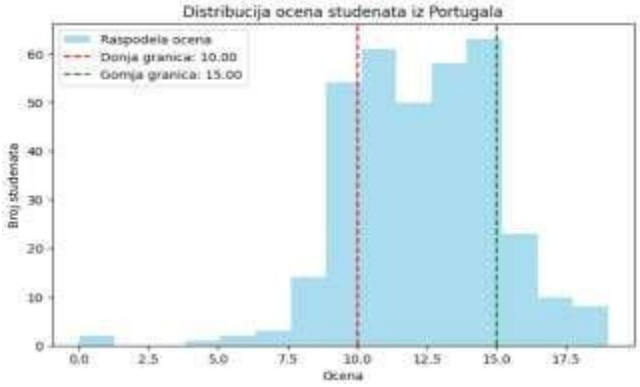
S obzirom da ovi atributi nisu jednako predstavljeni u skupovima podataka, sproveden je niz transformacija nad svakim, kako bi mogli da se spoje u jedan, a potom i porede vrednosti atributa po državama i vrši obuka modela. Nakon transformacija i spajanja dobijen je skup podataka sa podacima o 870 studenata. Ocene (uspeh studenata) su svedene na kategorije uspešan – A, srednje uspešan – B i slabog uspeha – C, jer su tako bile izražene u skupu podataka o studentima iz Pakistana, pa je logičan sled bio da brojčane vrednosti iz ostalih skupova svrstamo u kategorije.

III.I Priprema podataka

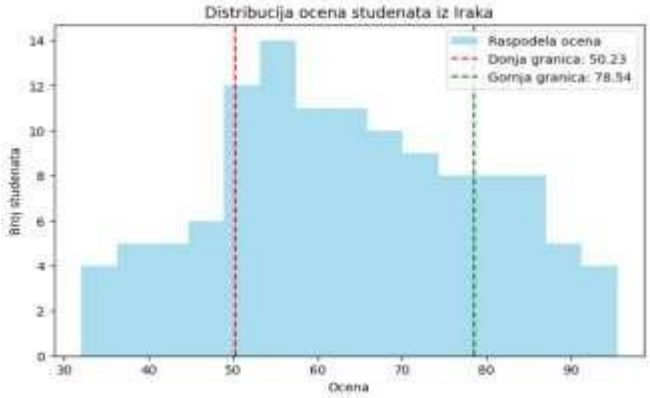
Priprema [4] skupa podataka počela je tako što su obrisani redovi koji nisu imali vrednost u koloni sa prosečnom ocenom (Matric/GPA), jer bez nje nije moguće vršiti klasifikaciju. Ocene su zatim pretvorene u numerički oblik kako bi mogle da se analiziraju i prikazuju na grafikonima. Histogrami i raspodela pokazali su da ocene nisu ravnomerno raspoređene i da postoje pojedini ekstremni slučajevi. Da bi se to ublažilo, definisane su granice pomoću kvantila i interkvartilnog raspona (IQR), pri čemu su izabrane vrednosti koje bolje hvataju sredinu raspodele. Umesto da se koristi stroga podela na 50% i 80% kao granice, kvantili su dali realističniju raspodelu, jer je inače klasa najboljih (A) bila premala. Tako su formirane tri kategorije: C za najslabije rezultate, B za srednje, i A za najbolje. Važno je napomenuti da ekstremne vrednosti nisu obrisane, već su posmatrane i korišćene pri kategorizaciji, kako se ne bi izgubilo previše podataka. Pored ocena, i ostale promenljive su obrađene. Pol je sveden na oznake „F" i „M", godina studija na brojeve (1, 2, 3…), a fakultet prepisan u posebno polje „oblast". Pošto nije bilo direktnih podataka o satima učenja, korišćena je aproksimacija na osnovu učestalosti izlazaka, pa je to polje prebačeno u procenjene sate učenja. Prisustvo na nastavi podeljeno je u tri grupe: vrlo dobro, dobro i loše, prema broju izostanaka. Smeštaj je klasifikovan na „Private" i „Non-private", dok je finansijski status određen na osnovu vrednosti mesečnog džeparca. Na kraju su granice za ocene vizuelno proverene grafikonima, a upoređeni su i rezultati sa apsolutnim pragovima (50/80%) i sa kvantilnim. Pokazalo se da kvantilna metoda daje ravnomerniju raspodelu i stabilnije rezultate za klasifikaciju.



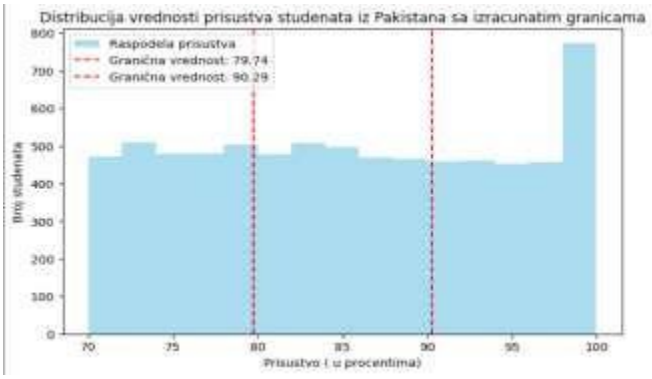
Kod [7] skupa podataka prvo su uklonjeni redovi sa praznim vrednostima u kolonama koje su bile ključne za analizu. Ocene su pretvorene u numerički oblik, a višestruke kolone koje opisuju različite testove i parcijale objedinjene su u jedinstven prosek. Prisustvo i ponašanje učenika (npr. alkohol, slobodno vreme, druženje) transformisani su u diskretne kategorije kako bi mogli da se porede sa sličnim atributima u drugim skupovima. Pošto su ocene bile izražene na skali od 0 do 20, urađena je standardizacija na procente (0–100%) kako bi bile uporedive sa afričkim i ostalim skupovima. Outlieri nisu uklanjani, već su posmatrani i ostavljeni u datasetu zbog relativno ograničenog broja učenika. Pol i godine su ostavljeni u izvornom obliku, dok je smeštaj sveden na binarne vrednosti. Finansijski status nije direktno postojao, pa su procene urađene na osnovu roditeljskog obrazovanja i zanimanja.



U [6] skupu podataka primenjen je postupak sličan postupku nad [4]. Prvo su obrisani redovi bez ocene, a zatim su ocene standardizovane u procente. Histogrami su pokazali nešto ravnomerniju raspodelu nego kod [4], ali i dalje sa izraženim ekstremima. Da bi se dobila logična klasifikacija, korišćeni su kvantili i IQR za određivanje granica. Na osnovu njih su formirane tri kategorije (A, B, C), s tim da je granica između B i A postavljena nešto niže nego što bi bila stroga vrednost od 80%, kako bi se postigla ravnoteža klasa. Pored ocena, dodatne transformacije su uključivale kodiranje pola, godina i smerova studija u numeričke ili binarne vrednosti. Prisustvo na nastavi mapirano je u kategorije. Za procenu sati učenja korišćene su postojeće kolone koje su opisivale učeničke navike, a u nedostatku tih podataka uvedene su aproksimacije. Finansijski status nije bio precizno definisan, pa su korišćene indikativne kolone (npr. posao roditelja) i prevedene u tri nivoa standarda.



[5] dataset je takođe prvo očišćen od redova bez ključnih vrednosti (ocena). Ocene su bile izražene u različitim formatima, pa su pretvorene u procente da bi se ujednačile sa ostalim skupovima. Nakon toga je primenjena standardizacija, a histogramska analiza otkrila je dosta koncentrisanih vrednosti u srednjem opsegu, dok su najbolji i najslabiji rezultati bili manje zastupljeni. Da bi se pravilno klasifikovali studenti, korišćeni su kvantili (Q1 i Q3) i IQR metod za definisanje granica. Outlieri nisu brisani, ali su posmatrani pri određivanju kategorija. Pol i godina studija su normalizovani u standardne vrednosti, dok su podaci o prisustvu i učenju aproksimirani i podeljeni u manje kategorije (npr. „slabo“, „srednje“, „dobro“). Finansijski status je kod pakistanskog skupa određen jednostavnije nego u [4], jer su postojale jasnije vrednosti u podacima, koje su direktno grupisane u tri nivoa.



III.II SPAJANJE PODATAKA

Nakon što su pojedinačni skupovi podataka objedinjeni u jedinstveni, sprovedene su dodatne transformacije kako bi se osigurala kvalitativna i kvantitativna konzistentnost podataka. U prvom koraku izvršena je analiza kompletiranosti i pouzdanosti podataka, pri čemu su identifikovane i uklonjene vrednosti koje bi mogle negativno uticati na performanse

modela. Konkretno, uklonjeni su nepotpuni zapisi sa nedostajućim ključnim atributima.

Posebna pažnja posvećena je [5], koji je zbog svoje strukture i specifičnih karakteristika pokazao potencijalno negativan uticaj na ukupnu tačnost modela. Naime, analiza i prethodna iskustva timova koji su radili sa ovim datasetom ukazala su na visoku raznolikost i neujednačenost podataka, što je moglo dovesti do smanjenja performansi prilikom kombinovanja sa ostalim datasetovima. Stoga je odlučeno da se koristi uzorak ovog skupa podataka, čime se zadržava reprezentativnost, ali se smanjuje negativan efekat na tačnost konačnog modela.

Po završetku čišćenja i transformacija, objedinjeni dataset je podeljen na trening i validacioni skup u sledećim proporcijama, u skladu sa dokumentovanim procedurama: trening skup**:** 80% ukupnog broja instanci, koji je korišćen za treniranje modela i optimizaciju parametara; validacioni/test skup**:** 10%-10% ukupnog broja instanci, korišćen za evaluaciju performansi modela i proveru generalizacije.

IV METOD

U okviru projekta cilj je bio da se razvije klasifikacioni model koji može da predvidi uspešnost učenika i studenata na osnovu podataka o njihovom životu i okruženju. Za potrebe istraživanja korišćeno je više javno dostupnih skupova podataka sa platforme *Kaggle*, koji su prikupljeni u različitim državama (Irak, Portugal, Pakistan i nekoliko afričkih zemalja). Da bi se dobio obuhvatniji i raznovrsniji skup podataka, datasetovi su objedinjeni na osnovu zajedničkih kolona. Nakon integracije, dobijeni su sledeći atributi: pol, godina studija, oblast (struka), država, sati učenja nedeljno, prisustvo na nastavi, smeštaj (privatni ili državni), finansijski status, bliskost sa roditeljima, edukacija roditelja i emotivni status (u vezi ili ne). Ciljna promenljiva bila je **uspešnost**, klasifikovana u tri kategorije:

- A – najuspešniji učenici,
- B – srednje uspešni,
- C – najmanje uspešni.

Za klasifikaciju su korišćene tri poznate metode:

- **K-Nearest Neighbors (KNN)**, jednostavan algoritam zasnovan na merenju udaljenosti između instanci.
- **Random Forest**, ansambl metoda bazirana na kombinaciji više stabala odlučivanja.
- **Multilayer Perceptron (MLP)** – neuronska mreža implementirana pomoću *scikit-learn* biblioteke.

Optimizacija hiperparametara sprovedena je uz pomoć *GridSearchCV* sa **unakrsnom validacijom od 5 podela (cv=5)**. Na ovaj način odabrane su vrednosti parametara poput broja suseda za KNN, broja stabala za Random Forest, i broja neurona u skrivenom sloju za MLP.

V REZULTATI I DISKUSIJA

Eksperimenti su sprovedeni podelom podataka na **trening, validacioni i test skup**. Kao mera performansi korišćena je **tačnost (accuracy)**, budući da je cilj bio da se oceni sposobnost modela da ispravno klasifikuje studente u tri grupe uspešnosti.

Dobijeni rezultati prikazani su u tabeli:

Model	Validacioni skup	Test skup
KNN	57%	60%
Random Forest	68%	54%
MLP (neuronska mreža)	62%	56%

Analiza rezultata pokazuje da **Random Forest** daje najbolje rezultate na validacionom skupu, ali se generalizacija na test skupu pokazala lošijom, što ukazuje na **problem prenaučenosti (overfitting)**. KNN i MLP su davali stabilnije rezultate, ali su ukupno imali nižu tačnost.

Glavni izazov bio je kvalitet podataka. Spojeni datasetovi sadržali su nedovoljno primera i nedostatke u popunjenosti, što je značajno ograničilo performanse modela. Takođe, različito poreklo podataka (različite zemlje, obrazovni sistemi i socijalni konteksti) uvelo je dodatni šum i otežalo učenje modela.

U literaturi se često ističe da je Random Forest robustan u radu sa heterogenim podacima, što se delimično poklopilo sa našim rezultatima na validacionom skupu. Međutim, slabija generalizacija na test skupu ukazuje da bi za ovakav zadatak bila potrebna dalja **obrada podataka (feature engineering, balansiranje klasa, uklanjanje šuma)** ili prikupljanje kvalitetnijih i konzistentnijih podataka.

Zaključujemo da iako nijedan od testiranih modela nije dostigao visoku tačnost, eksperiment pokazuje da je problem klasifikacije uspešnosti učenika kompleksan i da zahteva bogatije podatke i sofisticiranije metode obrade kako bi se dobili stabilniji i pouzdaniji rezultati.

VI ZAKLJUČAK

U ovom radu bavili smo se problemom predviđanja uspešnosti učenika i studenata na osnovu podataka o njihovom životu, učenju i socijalnom okruženju. Motivacija je bila da se razvije model koji može da generalizuje na različite populacije spajanjem više dostupnih datasetova iz različitih zemalja. Iako su korišćeni modeli (KNN, Random Forest i MLP) postigli umerene rezultate, istraživanje je pokazalo da kvalitet i konzistentnost podataka imaju ključnu ulogu u performansama klasifikacionih sistema. Glavna vrednost rada ogleda se u prikazu izazova i ograničenja pri radu sa heterogenim podacima, a potencijalne buduće smernice uključuju prikupljanje bogatijih i homogenijih skupova podataka, primenu naprednijih tehnika obrade i balansiranja klasa, kao i evaluaciju na širem spektru metoda, što bi doprinelo pouzdanijim predikcijama u obrazovnim istraživanjima.

LITERATURA

[1] A. S. Hashim, W. A. Awadh, A. K. Hamoud, College of Computer Science and Information Technology / University of Basrah, Basrah Iraq *Student Performance Prediction Model based on Supervised Machine Learning Algorithms*

[2] Indonesian Journal of Electrical Engineering and Computer Science ρ 1584 Vol. 16, No. 3, December 2019, pp. 1584~1592 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i3.pp1584-1592 *Supervised data mining approach for predicting student performance* Wan Fairos Wan Yaacob, Syerina Azlin Md Nasir, Wan Faizah Wan Yaacob, Norafefah Mohd Sobri Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Malaysia

[3] International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-4C, April 2019 *Educational Data Mining: Student Performance Prediction in Academic* Y. K. Salal, S. M. Abdullaev, Mukesh Kum

[4] Effects of Alcohol on Student Performance.

[5] Factors_ affecting_ university_student_grades

[6] Iraqi Student Performance Prediction

[7] Student Performance