Predikcija nivoa adaptacije studenata na online nastavu

Tamara Ranisović IN 1/2020, ranisovic123@gmail.com

Tijana Varađanin IN 31/2020, tijana.varadjanin@gmail.com

I. UVOD

Tema projekta je predikcija nivoa adaptacije studenata na online nastavu u zavisnosti od uslova u kojim je student pratio nastavu. Tokom pandemije koronavirusa, online nastava postala je ključni oblik obrazovanja širom sveta. Ova tranzicija predstavljala je izazov za sve činioce obrazovnog sistema, uključujući učenike, nastavnike i roditelje. Dok je online nastava pružila siguran način učenja tokom pandemije, istovremeno je naglasila važnost prilagodljivosti, tehnološke pismenosti i podrške kako bi se osiguralo uspešno učenje na daljinu.

Postoji nekoliko razloga zašto je važno istraživati ovu temu:

1. Rano upozorenje: Identifikacija studenata koji bi mogli imati poteškoća s online nastavom može omogućiti ranu intervenciju i pružanje dodatne podrške. Time se može smanjiti stopa odustajanja i povećati uspeh studenata.
2. Optimizacija resursa u obrazovnim institucijama: Ako se razumeju faktori i prepoznaju određeni obrasci adaptacije koji su povezani s određenim grupama studenata, resursi se mogu usmeriti na podršku tim grupama.
3. Kvalitet online nastave: Ako se prilikom analize i predikcije nivoa adaptacije primeti da se veći broj studenata ne snalazi nabolje, to može ukazati na potrebu za poboljšanjem metoda online nastave.
4. Istraživanje učinkovitosti adaptivnih strategija: Modeli za predikciju mogu se koristiti i u istraživačke svrhe kako bi se bolje razumeli učinci različitih metoda podrške ili tehnoloških alata na adaptaciju studenata.

Kreira se model za predikciju nivoa adaptacije studenata na online nastavu zasnovan na klasifikacionom problemu sa 3 klase (nizak, srednji i visok nivo), koji koristi različite algoritme mašinskog učenja: kNN, SVM i stabla odluke.

II. BAZA PODATAKA

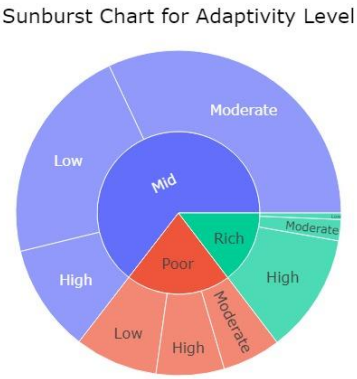
Pre četiri godine, učenici i studenti iz Bangladeša podelili su svoja iskustva sa praćenjem nastave od kuće kroz upitnik. Jedan uzorak iz baze predstavlja odgovor jednog studenta na anketu o obrazovanju i uslovima u kojim prati online nastavu. U bazi ima 1205 uzoraka i 14
obeležja, koja su sva kategorička. Obeležja su: pol, starosna grupa, nivo obrazovanja, tip obrazovne institucije,
IT student, lokacija, nivo isključenja struje usled opterećenja, finansijska situacija, tip interneta, tip mreže, dnevno trajanje nastavne, dostupnost institucionog LMS-a (Learning Management System), uređaj i nivo adaptacije.

III. METODOLOGIJA RADA

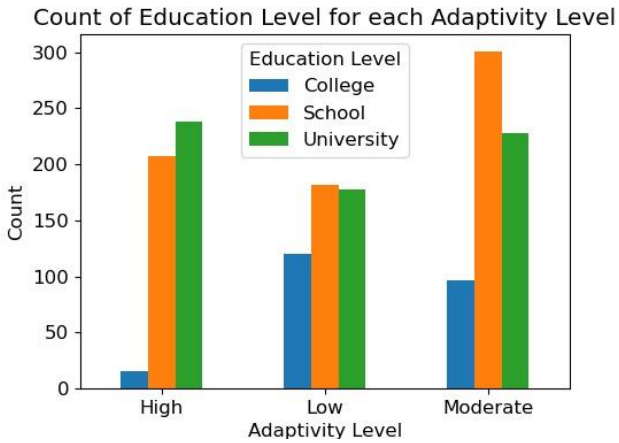
U ovoj sekciji rada sledi detaljan opis koraka sprovedenog istraživanja.

A. Eksplorativna analiza podataka

Kako u bazi nema nevalidnih niti nedostajućih vrednosti ni za jedno obeležje, nije bilo potrebe za čišćenjem podataka. Početni skup obeležja je u celosti relevantan. Dat je grafički prikaz adaptacije studenata različitih kategorija nekih obeležja (*Slike 1 i 2*).

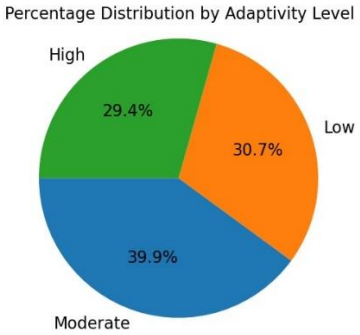


Slika 1. Zastupljenost nivoa adaptacije po finansijskim kategorijama



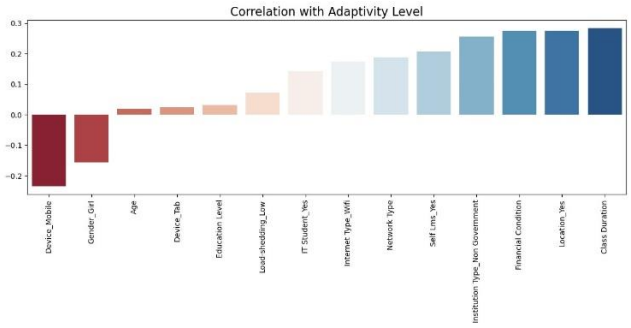
Slika 2. Zastupljenost kategorija nivoa obrazovanja po nivoima adaptacije

Ciljna promeljiva nivo adaptacije svrstava uzorke u jednu od tri kategorije: nizak, srednji i visok. Izvršena je podela skupa na trening i test skup u odnosu 90:10. Usled neravnomerne zastupljenosti uzoraka po kategorijama, izvršeno je balansiranje ciljne promenljive na trening skupu povećavanjem broja uzoraka najmanje zastupljene klase četiri puta (*High*) (*Slika 3*).



Slika 3. Balansirana raspodela uzoraka prema nivou adaptivnosti

Kategorička obeležja su pretvorena u numerička korišćenjem *dummy* varijabli za binarna obeležja i obeležje uređaj (D*evice)* kod kojeg ne postoji logički poredak kategorija. Ostala obeležja imaju kategorije koje se mogu rangirati, te su bile pogodne za *label encoding.* Nakon konverzije, prikazana je korelacija između svakog obeležja i ciljne promenljive *(Slika 4).*



Slika 4. Korelacija sa ciljnom promenljivom

B. Obuka modela

Važan korak pripreme skupa podataka koji može znatno olakšati obuku modela je redukcija dimenzionalnosti, sa ciljem identifikacije najinformativnijih komponenti, uz očuvanje što veće varijanse prisutne u podacima. LDA metoda nije pogodna zbog ograničenja broja komponenti na: $n\_components \leq min\{n\_classes - 1, n\_features\}$, gde je: n_components - broj komponenti; n_classes - broj klasa ciljne promenljive (3); n_features - broj obeležja bez ciljne promenljive (13).
Stoga je korišćena PCA metoda, koja se zasniva na kreiranju glavnih komponenti linearnom kombinacijom originalnih obeležja. Zarad očuvanja varijanse od 95%, bilo je potrebno izdvojiti 10 komponenti.

Za automatsko traženje optimalnih hiperparametara modela upotrebljena je GridSearchCV, tehnika koja prolazi kroz sve moguće kombinacije istih, koristeći unakrsnu validaciju za procenu performansi modela. Pri tom procesu, trening skup je podeljen na 5 podskupova, čime se postiže razumna ravnoteža između dobijanja pouzdane procene performansi i korišćenja optimalnih računarskih resursa. Za vrednovanje performansi modela ovog problema najrelevantnija mera uspešnosti je tačnost (accuracy), koja predstavlja odnos broja ispravno klasifikovanih uzoraka i ukupnog broja uzoraka.

IV. MODELI

U ovom poglavlju biće predstavljeni odabrani i ujedno najčešće korišćeni algoritmi - kNN, SVM i stabla odluke, njihove ideje kao i primena za ovaj problem klasifikacije.

1. *kNN*

K-najbližih komšija (KNN) je jednostavan, nadgledani algoritam mašinskog učenja koji se koristi za klasifikaciju i regresiju. Ključna ideja na kojoj je algoritam zasnovan je da kada dobije novi, nepoznati uzorak za klasifikaciju, računa njegovu udaljenost od ostalih uzoraka u trening skupu. Zatim, algoritam bira K najbližih susednih uzorak

iz trening skupa na osnovu izračunatih udaljenosti. Na kraju, novi uzorak se klasifikuje kao pripadnik klase koja je najčešća među K suseda.

Vrednost parametra K i različite metrike rastojanja imaju značajan uticaj na performanse modela. U ovom slučaju, korišćeni hiperparametri za GridSearchCV su:

- *n_neighbors*: predstavlja broj najbližih suseda koji su uzeti u obzir. Testirane su vrednosti od 1 do 21, sa korakom 2.
- *weights*: parametar koji određuje da li će udaljenosti biti ravnomerno tretirane (*uniform*) ili će bliži susedi imati veći uticaj (*distance*).
- *metric*: uloga ovog parametra je određivanje tipa metrike koji se koristi za računanje udaljenosti između tačaka. Testirane su četiri različite: *jaccard, hamming, dice* i *euclidean*.

Nakon završetka pretrage, najbolji parametri koji se koriste za konačni model su: {*metric: euclidean, n_neighbors:* 11, *weights: distance*}.

2. *SVM*

SVM (Support Vector Machine) je algoritam mašinskog učenja za klasifikaciju koji se temelji na pronalaženju hiperravni koja najbolje razdvaja instance različitih klasa u višedimenzionalnom prostoru karakteristika. Hiperravan nastoji da maksimizira rastojanje između najbližih tačaka različitih klasa. Dimenzija hiperravni zavisi od broja obeležja. Za slučaj višeklasne klasifikacije koriste se strategije kao što su jedan protiv svih (one-vs-rest) ili jedan protiv jednog (one-vs-one). Kod prvog pristupa kreira se po jedan klasifikator za svaku klasu, pri čemu se vrši nadmetanje sa svim ostalim, a klasa sa najvišom verovatnoćom se smatra klasom konačne klasifikacije. U drugom slučaju, klasifikatori se porede po parovima, te novi uzorak dobija pripadnost onoj klasi koja je najveći broj puta "izašla kao pobednik".

Kako bi se postigla što veća tačnost modela , izvršeno je variranje sledećih hiperparametara:

- Regularizacioni parametar C: uspostavlja ravnotežu između postizanja što manje greške na trening skupu i sprečavanja prenaučenosti. Povećavanjem njegove vrednosti, model teži ka minimizaciji greške na trening skupu, što rezultuje složenijim modelom i eventualnim natprilagođenjem. Obrnuto, smanjivanje vrednosti C doprinosi povećanju tolerancije na greške, pri čemu može doći do veće sposobnosti generalizacije modela, ali uz moguće smanjenje preciznosti na trening skupu. Testirane su vrednosti koje pokrivaju različite redove veličina (0.1, 1, 10, 100 i 1000), što omogućava istraživanje širokog spektra regularizacionih snaga i uticaja na performanse modela. Vrednosti poput odabranih često su korišćene u praksi i smatraju se standardnim izborom za početak optimizacije parametra C.
- *gamma*: određuje koeficijent jezgra u modelu. Velika vrednost gamma dovodi do uskog granica odluke, što može dovesti do prenaučenosti, dok manja vrednost gamma omogućava granicama odluke da budu šire, ali može dovesti do podnaučenosti. Ovde su odabrane vrednosti 0.1, 0.01, 0.001, i 0.0001 kako bi se istražili različiti efekti gamma parametra.
- kernel: određuje vrstu jezgra koje se koristi u SVM modelu. Različite vrste jezgara mogu imati različite efekte na sposobnost modela da se prilagodi podacima. Ovde su korišćene tri vrste jezgra: RBF (radijalna bazna funkcija), polinomijalno i sigmoidno.
- *decision_function_shape*: određuje strategiju za višeklasnu klasifikaciju. *Ovo* označava strategiju jedan protiv jednog (one-vs-one), dok *ovr* označava strategiju jedan protiv svih (one-vs-rest).

Na kraju, najbolji parametri koji se koriste za konačni model su:
{C: 0.01, decision_function_shape: ovo, gamma: 100, kernel: poly}.

3. *Stabla odluke*

Obuka modela predstavlja formiranje stabla sa korenskim čvorom koji sadrži ceo trening skup podataka. Zatim se bira obeležje koje deli podatke na što čistije čvorove, na osnovu nekog kriterijuma, kao što su Đinijev indeks i entropija. Proces se rekurzivno ponavlja koristeći top-down greedy podelu, tako što deli čvorove ne obazirući se na mogućnost bolje podele u budućnosti. Ovaj proces se nastavlja sve dok svi listovi ne sadrže podatke iste klase ili dok ne bude zadovoljen neki kriterijum zaustavljanja. Moguće je pustiti stablo da se razgrana, te primeniti postupak *pruninga* (potkresivanja), koji spaja čvorove ili uklanja delove stabla koji ne doprinose poboljšanju tačnosti na validacionom skupu. Na listovima stabla nalaze se klase, čije vrednosti putem većinskog pristupa određuju klasu datog uzorka. Prednost ovog algoritma ogleda se u

jednostavnosti i maloj zahtevnosti određivanja obeležja i zadavanja kriterijuma. Dobijeni klasifikator je kompaktan i brz, nezavistan od tipa podataka i lak za interpretaciju. Takođe, ima dobru izdržljivost u odnosu na uzorke koji odstupaju od populacije. Međutim, fragmentacija podataka može uticati negativno na sam proces, jer se vremenom odluke vrše na osnovu sve manje i manje podataka. U odnosu na SVM i druge algoritme, nema naročito bolje performanse, te se zato najviše koristi u ansamblima klasifikatora.

Da bi se postigla što veća tačnost modela , izvršeno je testiranje sledećih hiperparametara:

- *max_depth*: označava maksimalnu dubinu stabla odlučivanja. Zadavanje opsega omogućava algoritmu da testira različite dubine i pronađe optimalnu. Dublje stablo može dovesti do natprilagođenja, dok pliće stablo može rezultirati podnaučenosti.
- *min_samples_leaf*: predstavlja minimalni broj uzoraka u lisnom čvoru. Kontrolisanjem broja uzoraka u listovima, algoritam istražuje različite konfiguracije kako bi sprečio prenaučenost.
- *min_samples_split*: parametar koji predstavlja kriterijum podele čvora, odnosno minimalni broj uzoraka potreban za taj proces.
- *criterion*: mere nečistoće *gini* i *entropy* su kriterijumi za procenu kvaliteta podele čvora. Cilj je da budu što manje, jer tada rezultuju čistijim, više homogenim čvorovima.
- splitter: određuje strategiju (*best*-najbolje ili *random*-nasumično) koja se koristi za odabir podele na svakom čvoru.
- max_features: predstavlja broj obeležja koji će stablo razmatrati pri izboru najboljeg pitanja za podelu čvora. Vrednosti testiranih parametara su: *log2, sqrt, auto*.

Optimalni parametri su:
{'criterion': 'entropy', 'max_depth': 13, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 3, 'splitter': 'best'}

V. REZULTATI I ZAKLJUČAK

Pronađeni optimalni hiperparametri za sva tri algoritma uvršteni su u kreiranje odgovarajućih konačnih modela. U ovoj sekciji prikazani su rezultati svakog klasifikatora na test skupu. U tabeli ispod prikazane su vrednosti mera uspešnosti: tačnost (*accuracy*), preciznost (*precision*), osetljivost (*recall*) i F-mera (*f1-score*). Mera tačnosti uzeta je kao krucijalna u određivanju performansi obrađenih modela.

Dok se neki student uspešno savladali izazove i prilagodili se novom okruženju, drugi su se suočili s poteškoćama. Ova raznolikost može biti rezultat različitih faktora, uključujući tehničke sposobnosti, pristup tehnologiji, motivacija, podrška porodice i okoline, kao i specifične karakteristike nastavnih programa. Istraživanje naglašava potrebu za daljim istraživanjem kako bi se bolje razumele individualne potrebe i strategije adaptacije studenata te kako bi se osmislili prilagođeni pedagoški pristupi koji bi podsticali njihovu uspešnost u online nastavi.

Model	Nivo adaptacije	Tačnost	Preciznost	Osetljivost	F-mera
kNN	nizak	92.56%	0.94	0.94	0.94
	srednji		0.95	0.90	0.93
	visok		0.77	1.00	0.87
SVM	nizak	90.91%	0.94	0.92	0.93
	srednji		0.92	0.90	0.91
	visok		0.75	0.90	0.82
Stabla odluke	nizak	91.74%	0.94	0.92	0.93
	srednji		0.93	0.90	0.92
	visok		0.77	1.00	0.87

Tabela 1. Prikaz mera uspešnosti konačnih modela

VI. LITERATURA

1. https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html
2. https://www.kaggle.com/datasets/mdmahmudulhasansuzan/students-adaptability-level-in-online-education
3. https://www.datacamp.com/blog/classification-machine-learning