Algoritam potpornih vektora u predviđanju uspjeha studenata

Marina Milićević
Fakultet za proizvodnju i mneadžment Trebinje
Univerzitet u Istočnom Sarajevu
BiH, Republika Srpska
marina.milicevic@fpm.ues.rs.ba

Sažetak—**U eri u kojoj tehnologija svojim razvojem neprestano preoblikuje polje profesionalne prakse, jasno je da je obrazovanje inženjera danas izazovan zadatak. U prilog tome govore i sve popularnije nastavne metode u kojima se tradicionalni načini učenja kombinuju sa modernim tehnologijama, a sve u cilju veće motivacije za učenje i rad kod samih učenika. Predmet istraživanja ovog rada je upravo jedna nastavna metoda, prilagodjena za nastavu na tehničkim fakultetima, a koja u svojoj pozadini koristi metode mašinskog učenja (metod potpornih vektora). Uvedena metoda prepoznaje značaj procjene mogućnosti i znanja studenata jer dobra klasifikacija studenata omogućava nastavniku da lakše koncipira sam čas i odabere pristup za koji smatra da će dati najbolje rezultate.**

Ključne riječi-metod potpornih vektora; algoritmi mašinskog učenja; klasifikacija studenata

I. UVOD

Jedan od izazova sa kojim se susreću sve organizacije današnjice, pa tako i univerziteti, jeste usvajanje inovacija. Posebno je interesantna situacija kada se govori o tehnološkim inovacijama u nastavnom procesu koje, iako zahtijevne, imaju širok spektar dobrobiti kako za same univerzitete, tako i za studente. Uključivanjem savremenih metoda, kao na primjer vještačke inteligencije i mašinskog učenja, u tradicionalne procese na univerzitetu može se doći do veoma korisnih informacija koje će usmjeriti akcije univerziteta u pravcu postizanja zadatih ciljeva.

Predvidjanje akademskog uspjeha studenata i njihova klasifikacija je uopšteno govoreći, skup metoda koje se primjenjuju za prikupljanje podataka i dobijanje informacija o akademskom uspjehu studenata, a koje najčešće uključuju neki od algoritama mašinskog učenja. Sve veća informatizacija univerziteta dovela je do toga da danas na univerzitetima postoje ogromne količine podataka o studentima. Unutar tih podataka skrivene su mnoge važne informacije za univerzitete, a pribavljanje tih informacija nije ni malo lak zadatak i neminovno uključuje metode vještačke inteligencije i mašinskog učenja. Neke od korisnih informacija mogu biti zavisnost akademskog uspjeha studenta od prosjeka ocjena u srednjoj školi ili stuktura studenata koji postižu najbolje (najlošije) rezultate. Što bolja iskoristljivost dostupnih podataka dovodi do boljeg znanja i akademskog uspjeha studenata, što u konačnici rezultuje boljom slikom i pozicijom samog univerziteta. Upravo su svakodnevni napori da se ljestvica kvaliteta u obrazovanju podigne na viši novo doveli do intenzivnijeg korišćenja algoritama mašinskog učenja prilikom rudarenja podataka na univerzitetima (EDM – Educational Data Mining).

Postoji veliki broj različitih algoritama mašinskog učenja razvijenih i prilagođenih za specifične zadatke, a zajedničko im je da svi počivaju na teoriji funkcionalne analize i statistike. Postojanje podataka i obrasca u njima su dva glavna preduslova za primjenu mašinskog učenja, a samo učenje možemo podijeliti na učenje sa nadgledanjem i bez nadgledanja. Prva kategorija se pokazala kao ona koja daje bolje rezultate, najviše zbog toga što su kod ove metode ulazni podaci označeni tj. pored skupa ulaznih podataka imamo i skup oznaka koje nam predstavljaju izlazne vrijednosti. Kod učenja bez nadgledanja nije unaprijed poznato kako će izgledati izlaz iz sistema jer radimo sa neoznačenim podacima. Osim ove dvije grupe, u praksi se često sreće i pojačano mašinsko učenje – posebna vrsta mašinskog učenja u kojoj agent uči iz interakcije sa okolinom metodom pokušaja i greške.

Algoritmi mašinskog učenja sa nadgledanjem mogu imati različite izlazne vrijednosti, pa ih možemo svrstati u dvije grupe: (1) algoritme za klasifikaciju koji na izlazu imaju vrijednosti iz skupa sa konačnim brojem elemenata i svaka instanca problema koji se posmatra će na izlazu uzeti jednu od ovih vrijednosti, odnosno svrstaće se u jednu od izlaznih klasa i (2) algoritme za regresiju koji na izlazu imaju samo jedan realan broj, a cilj im je na osnovu ulaznih vrijednosti naći funkciju koja će dati odgovarajuće izlazne vrijednosti.

Sa ciljem određivanja uspješnosti izabranog agoritma koriste se razne metrike, kao na primjer *srednja kvadratna greška* kod algoritama za regresiju. Kod algoritama za klasifikaciju najčešće se koristi *tačnost* - procenat tačno klasifikovanih podataka. U situacija kod kojih je broj primjera jedne klase znatno veći od broja primjera druge klase (neizbalansirani podaci) kod klasifikacijskih algoritama mašinskog učenja koristimo metrike kao što su F1 ocjena, kros-entropija i površina ispod ROC krive.

Osnovni problem mašinskog učenja je pretreniranje. Situacije kada rezultati algoritma mašinskog učenja nisu zadovoljavajući rješavamo dodavanjem novih karakteristika i

kreiranjem kompleksnije hipoteze pokušava se dobiti poboljšanje tokom treniranja algoritma. Međutim, postoji opasnost od pretreniranja (engl. overfitting) jer kompleksne funkcije mogu lako zabilježiti preslikavanja između ulaznih i izlaznih podataka tokom treniranja, ali usput i propustiti uočiti korisne obrasce i zakonitosti među podacima koji bi se zatim mogli primjeniti i u opštem slučaju (na potpuno novim podacima), a ne samo na podacima iz skupa za treniranje algoritma. Ako nam je pak funkcija suviše jednostavna, zapadamo u suprotan problem (podtrenitanje – engl. underfitting), jer jednostavne funkcije ne mogu da nauče kompleksne veze između ulaznih i izlaznih podataka.

U ovom radu predstavljen je model mašinskog učenja za predviđanje akademskog uspjeha studenta na jednom predmetu. Podaci o akademskom uspjehu studenata prve godine prikupljeni su na studijskim programima Industrijski menadžment i Industrijsko inženjerstvo za energetiku na Univerzitetu u Istočnom Sarajevu tokom perioda od 4 godine (od 2017.- do 2021.), a kao sredstvo za prikupljane podataka korišćen je informacioni sistem fakulteta. Kako su u pitanju studijski program tehničkog profila, predmet sa prve godine studija koji je izvojen kao najznačajniji i od čijeg savladavanja zavisi dalji tok studiranja za pojedinog studenta je matematika. Osnovni cilj istraživanja je uspjeti procjeniti na sredini semestra da li će student na kraju semestra položiti ili ne matematiku, pri tome koristeći podatke iz informacionog sistema univerziteta. Kako je matematika bazna disciplina koja prožima i spaja mnoge druge oblasti, a posebno oblasti tehničkog usmjerenja, polazi se od pretpostavke da se podaci o uspjehu studenta na predmetima tehničke i praktične orijentacije mogu iskoristit za predviđanje uspjeha studenta iz predmeta matematika. Kao podaci od značaja izdvojeni su: prosjek ocjena iz srednje škole, uspjeh na laboratorijskim vježbama tokom semestra i prosječan broj bodova iz tehničkih disciplina na sredini semestra. Algoritam mašinskog učenja koji je korišćen u radu je metod potpornih vektora (engl. Support Vector Machine – SVM), a kako smo rješavali klasifikacijski problem, dvije definisane klase označesne su sa F (student neće položiti predmet) i P (student će položiti predmet).

II. PREGLED LITERATURE

Sa brzim razvojem tehnologije i sve većim brojem podataka koji se prikupljaju na univerzitetima, različiti metodi mašinskog učenja su počeli da se intenzivno primjenljuju za predviđanje akademskog uspheja studenata. Međutim, razni istraživali koristili su različite tehnike i atribute (akademske, demografkse, sociološke itd) za predviđanje i klasifikaciju. U nastavku je dat kratak prikaz nekih od studija koje su od značaja za istraživače dato u ovom radu.

Pouzdan pregled za buduće istraživače na temu primjene vještačkih neuronskih mreža prilikom analize i prikupljanja podataka o studentima na univerzitetima (EDM) dat je u [1]. Pažnju posvećuju obradi i prikupljanju podataka, a zatim analiziraju na koji način autori koriste neuronske mreže u EDM svrhe. Najčešći tipovi neuronskih mreža su rekurentne neuronske mreže, zatim jednoslojne i višeslojne nepovratne (engl. feedforward) neuronske mreže, dok je učenje sa nadgledanjem najčešći tip algoritama koji se koristi pri čemu je srednja jvadratna greška pojavljuje kao najčešći metod za procjenu kvaliteta mreže, odnosno za cost functiom minimisation.

Stabla odlučivanja (engl. Decision Trees) su model mašinskog učenja koji se može primjeniti i za zadatke klasifikacije i za zadatke regresije. Autori studije [2] prikazali su dva modela tipa stabla odlučivanja u kojim su od većeg broja atributa za svakog studenta izabrana samo ona koja značajno utiču na predviđanje i klasifikaciju, pri čemu je varirana zavisna promjenljiva u stablima. U jednom modelu kao zavisna promjenjiva korišćena je prosječna ocjena, a u drugom vrijeme potrebno za završetak studija.

U radu [3] stablo odlučivanja je korišćeno za klasifikaciju studenata u zavisnosti od njihovog uspjeha na kursevima matematike, pri čemu su studenti razvrstani u tri klase: pass – studenti koji polažu predmet, fail – studenti koji neće položiti predmet i conditional – studenti koji su na granici između ove dvije klase. Eksperimentalni rezultati su pokazali da je tačnost modela 72%. Još jedna studija o primjeni stabla odlučivanja za klasifikaciju učenika na dvije klase u zavisnosti od stepena uspješnosti izrade konkretnog 3D modela data je u radu [4].

Jedan od najpopularnijih algoritama klasičnog mašinskog učenje je metod potpornih vektora SVM koji je usvojoj osnovi neprobabilistički binarni klasifikator. Posebno dobro se pokazao u problemima klasifikacije studenata. U radu [5] ovaj metod je korišćen za predviđanje da li će student napustiti ili završiti fakultet, dok je u istraživanje [6] SVM metod korišćen za predviđanje uspjeha studenata koristeći podatke dostupne sa društvenih mreža. Važno je napomenuti da primjenljivost algoritama mašinskog učenja mnogo široka i obuhvata discipline od energetike, socioloških nauka pa sve do izdvajanja semantike i obrade teksta [7], [8] i [9].

III. METODOLOGIJA

A. Metod potpornih vektora

Metod potpornih vektora (SVM) je vrsta algoritma koji se najviše koristi za klasifikacijske probleme, mada se može koristiti i za probleme regresije. Metod potpornih vektora je neprobabilistički (izlaz iz modela je samo odluka), binarni (podatke dijeli u jednu od dvije klase), linearni (proširenja su moguća) klasifikator. Cilj SVM metode je da se pronađe optimalna *hiperravan* koja maksimalno odvaja tačke podataka jedne klase od druge, odnosno koja ima maksimalnu marginu. Dimenzija hiperravni zavisi od broja obilježja: ako je broj obilježja jednak dva onda je hiperravan prava, dok je za tri obilježja hiperravan ravan dimenzije 2. Hiperravan se određuje kao maksimalna udaljenost podataka od granice odlučivanja – na taj način na određivanje granice odlučivanja uticaj imaju samo oni podaci koji su joj najbliži (*potporni vektori*), dok oni koji su joj daleko nemaju nikakav uticaj na proces optimizacije. U slučaju linearno odvojive binarne klasifikacije model predstavlja jednačinu hiperravni razdvajanja. Ako je n broj obilježja koji se koriste u modelu, onda je hipoteza oblika:

$$h(x) = W^T X + b = 0,$$

gdje je sa b označen *bias*, a W je *vektor normale* na hiperravan. Ukoliko se primjeri jedne klase označe sa $y = 1$, a primjeri druge klase sa $y = -1$, jednačine hiperravni koje naliježu na ovako definisane klase su:

$$W^T X + b = 1 \text{ i } W^T X + b = -1.$$

Širina ovako definisanih margina data je formulom $\frac{2}{\|W\|}$, pa se problem optimizacije sada definiše kao maksimizacija izraza $\frac{2}{\|W\|}$, odnosno minimizacija $\|W\|$.

Insistiranje na linearnoj separabilnosti klasa može da dovede do pretjerane prilagođenosti modela podacima, pa se umjesto toga dozvoljava ulazak podataka u marginu ili njegov prelazak na pogrešnu strani hiperravni (naravno, u određenoj mjeri). Margina u kojoj je ovo dozvoljeno naziva se *mekom marginom* (engl. soft margin).

B. *Podaci*

Prikupljanje i priprema podataka je bio prvi korak prilikom konstruisanja modela. Posmatrano je 50 studenata upisanih na prvu godinu industrijskog inženjerstva tokom 4 godine. Priprema podataka obuhvatala je eliminaciju nepotrebnih atributa (lični podaci, eliminacija uspjeha na opštim kursevima kao na primjer strani jezik). Podaci su podjeljeni u dvije grupe i to podaci za treniranje modela (80%) i podaci za testiranje modela (20%). Izdvojeno je 3 atributa kao potencionalno značajni za cilj istraživanja i to: prosjek ocjena iz srednje škole (GPA), uspješnost tokom laboratorijskih vježbi u toku semestra (LabExercises) i uspjeh u tehničkim disciplinama sa prve godine na sredini semestra (midterms). Tokom eksperimenta korišćena su samo posljednja dva atributa, dok je GPA zbog nedoslijednosti usljed različitih profila srednjih škola izostavljen iz modela. U tabeli I data je analiza izabranih atributa za model sa domenom za svaki atribut. Zavisni atribut tj. koji predviđamo je uspjeh studenta na kursu iz matematike koja se polaze na kraju akademske godine.

TABELA I. ZNAČAJNI ATRIBUTI SA OPISOM

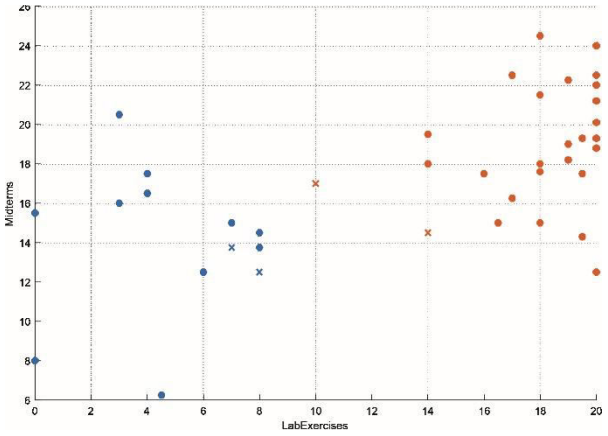
Atribut	Opis	Min	Max	Mean	Domen
GPA	Prosječan uspjeh iz srednje škole	2.50	4.78	3.61	Od 2.5 do 5.00
Lab exercises	Uspjeh iz labaratorijskih vježbi	0.00	20.00	13.89	Broj od 0 do 20
Midterms	Prosječna broj bodova iz tehničkih dispicplina na sredini semsetra	6.25	24.5	17.85	Broj od 0 do 25
Zavisni atribut	Uspjeh na kursu iz matematike	-	-	-	F, P

IV. REZULTATI

Model iz ove studije je razvijen i elsperiment je sproveden koristeći MATLAB programsko okruženje. Skup od 50 instanci prikupljenih podataka podjeljen je u odnosu 80:20 za treniranje i testiranje. Prilikom binarne klasifikacije koju smo sprovodili, zavisna promljenljiva "uspjeh na posmatranom predmetu" koja se u osnovi izražava kroz osvojen broj bodova je konvertovana u dvije klase: za broj bodova veći od 60 klasa P i za broj bodova manji od 60 klasa F.

Eksperimentalni rezultati su pokazali da je tačnost modela i na podacima za validaciju i na podacima za testiranje modela veoma visoka - 90%. Na slici 1 je prikazan tačkasti dijagram podataka u odnosu na dva posmatrana obilježja, pri čemu:

- crvena tačkica predstavlja tačno klasifikovanu klasu P;
- crveni znak x predstavlja pogrešno klasifikovanu klasu P;
- plava tačkica predstavlja tačno klasifikovanu klasu F;
- plavi znak x predstavlja pogrešno klasifikovanu klasu F.



Slika 1. Tačkasti dijagram (Scatter plot) podataka iz modela

Iz matrice zabune za trening podatke prikazane na tabeli II (lijevo) vidimo da od ukupno 27 instanci koje pripadaju klasi P, njih 25 je tačno klasifikovano, dok je njih 2 pogrešno klasifikovano kao klasa F. Od 13 instanci koje pripadaju klasi F, njih 11 je tačno klasifikovano, dok je njih 2 pogrešno klasifikovano kao klasa P. Sličan odnos tačno i pogrešno klasifikovanih instanci važi i za podatke za testiranje (tabela II desno).

TABELA II. MATRICA ZABUNE NA TRENING (LIJEVO) I TEST PODACIMA (DESNO)

Predviđena klasa

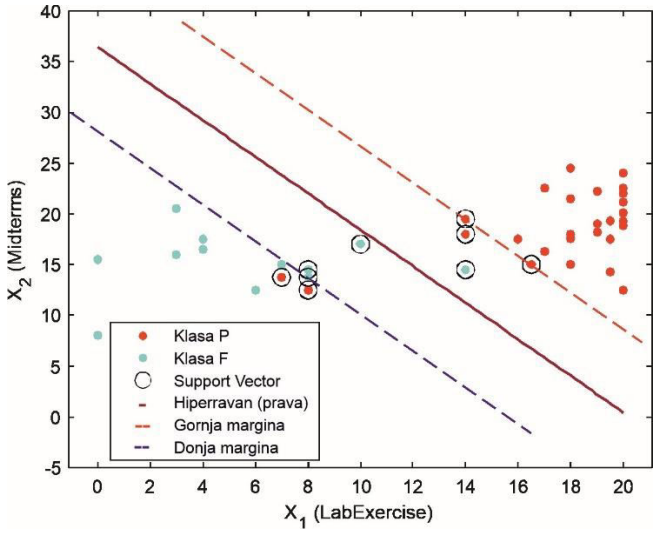
Stvarna klasa	F	P
F	11	2
P	2	25

Predviđena klasa

Stvarna klasa	F	P
F	1	1
P	0	8

Na slici 2 prikazana je hiperravan razdvajanja sa gornjom i donjom mekom marginom i potpornim vektorima.

Za ovako definisanu hiperravan imamo sljedeće karakteristike: bias: *b=-4.4020* i linearni koeficijenti predviđanja (beta): *0.2175* i *0.1209*.



Slika 2. Hiperravan razdvajanja sa potpornim vektorima mekim marginama

V. ZAKLJUČAK

Rezultat istraživanje u radu je predloženi model za klasifikaciju studenata prve godine tehničkih fakulteta u zavisnosti od njihovog uspjeha na predmetu matematika kao jedne osnovne discipline. Model koji je u radu predstavlje je jedan od najpopularnijih algoritama mašinskog učenja – metod potpornih vektora. Na akademski uspjeh studenata utiču razni faktori, ali u ovom radu smo koristili podatke dostupne iz informacionog sistema univerziteta.

Na osnovu istraživanja sprovedenog u radu možemo sumirati sljedeće zaključke:

(1) Model potpornih vektora predstavljen u radu može se koristiti za klasifikaciju studenata u dvije klasa na osnovu predviđenog uspjeha studenta na posmatranom predmetu.

(2) Studija je takođe dala uvid u činjenicu da se podaci o studentima pothranjeni u informacionim sistemima univerziteta mogu koristiti za predviđanje akademskog uspjeha studenata.

(3) Kao što je to pokazano u modelu, postoji jaka povezanost između uspjeha studenata na predmetima tehničke struke i uspjeha studenta na predmetu matematika.

Istraživanje predstavljeno u radu može sa proširi u nekoliko pravaca. Najprije, moguće je povećati dimenziju i posmatrati veći broj atributa. Osim toga, razni algoritmi mašinskog učenja se mogu kombinovati kako bi se dobio jedan hibridni model za klasifikaciju i predviđanje uspjeha studenata.

LITERATURA

[1] E. Okewu, P. Adewole, S. Misra, R. Maskeliunas, and R. Damasevicius, "Artificial neural networks for educational data mining in higher education: a systematic literature review," Appl. Artif. Intell. 35 (2021), no. 13, 983–1021.

[2] M. Dragičević, M. P. Bach, and V. Simičević, "Improving university operations with data mining: predicting student performance," Int. J. Econ. Manag. Eng. 8 (2014), no. 4, 1101–1106.

[3] M. Milićević, B. Marinović, and Lj. Jeftić, "Machine learning methods as auxiliary tool for effective mathematics teaching," Comput. Appl. Eng. Educ. (2024), e22787.

[4] M. Miljanović, M. Milićević, M. Jokanović Đajić, and S. Dostinić, "FabLab: Encouraging the Creativity of Secondary School Students of Various Professional Orientations," Revista De Education, Spain, 2024.

[5] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, "Predicting Student Performance using Advanced Learning Analytics," In Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion), Republic and Canton of Geneva, 2017.

[6] Benablo, C. I. P., Sarte, E. T., Dormido, J. M. D., and Palaoag, T. "Higher education Student's academic performance analysis through predictive analytics," In Proceedings of the 2018 7th International Conference on Software and Computer Applications—ICSCA 2018. New York.

[7] M. Milicevic and B. Marinovic, "Machine learning methods in forecasting solar photovoltaic energy production," Therm. Sci. 28 (2024), 479–488.

[8] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," J. Inform. Sci. 44 (2018), no. 1, 28–47.

[9] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," Expert Syst. Appl. 57 (2016), 232–247.

ABSTRACT

In an era where technology with its development is constantly reshaping the field of professional practice, it is clear that the education of engineers today is a challenging task. This is supported by the increasingly popular teaching methods in which traditional learning methods are combined with modern technologies, all with the aim of greater motivation for learning and work among the students themselves. The subject of research in this paper is one teaching method, adapted for teaching at technical faculties, which uses machine learning methods (support vector method) in its background. The introduced method recognizes the importance of assessing students' abilities and knowledge because a good classification of students allows the teacher to more easily design the lesson itself and choose the approach that he believes will give the best results.

Keywords: Suport vector machine algorithm; machine learning algorithms; students' classification

SUPPORT VECTOR MACHINE ALGORITHM IN PREDICTING STUDENT SUCCESS

Marina Milićević