Usporedba tehnika strojnog učenja u predikciji akademske uspješnosti

Grgić, Nikola

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Nikola Grgić

USPOREDBA TEHNIKA STROJNOG UČENJA U PREDIKCIJI AKADEMSKE USPJEŠNOSTI

ZAVRŠNI RAD

Varaždin, 2023.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Nikola Grgić

Matični broj: 35918/07–R

Studij: Informacijski sustavi i poslovni sustavi

USPOREDBA TEHNIKA STROJNOG UČENJA U PREDIKCIJI AKADEMSKE USPJEŠNOSTI

ZAVRŠNI RAD

Mentor/Mentorica:

Izv. prof. doc. dr. sc. Dijana Oreški

Varaždin, rujan 2023.

Nikola Grgić

Izjava o izvornosti

Izjavljujem da je moj završni/diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Ovaj rad u fokusu ima primjenu strojnog učenja za predikciju akademske uspješnosti. Bit će korištene tri tehnike strojnog učenja: stabla odlučivanja, metoda k-najbližih susjeda i neuronske mreže. Cilj rada je međusobno usporediti tehnike strojnog učenja te otkriti nove povezanosti između osobina studenta/studentice i akademske uspješnosti. Rad se bavi osnovama strojnog učenja, ali je potrebno osnovno predznanje iz matematike i informatike. U početku će biti govora o teorijskim osnovama strojnog učenja i svake od navedenih metoda, a potom će se objasniti praktični problem predikcije akademske uspješnosti te implementacija navedenih tehnika strojnog učenja u svrhu rješavanja tog problema. Na kraju će se analizirati rezultati te doći do zaključka koja je metoda pokazala najbolje performanse te ima li potencijala za primjenu u stvarnom okruženju.

Ključne riječi: akademska uspješnost, umjetna inteligencija, predikcija, usporedba performansi, neuronske mreže, K -najbližih susjeda, stabla odlučivanja

Sadržaj

1. Uvod

U Republici Hrvatskoj je u 2022. godini bilo približno 150 tisuća studenata (Državni zavod za statistiku, 2022). Jasno je da je studentima i obrazovnim ustanovama u cilju primijeniti suvremene tehnologije kako bi poboljšali akademsku uspješnost. U tom kontekstu, strojno učenje bi moglo predstavljati važnu komponentu za analizu i predviđanje metrika koje signaliziraju akademske performanse studenta. Na ovaj način moguće je analizirati svačiju osobnu situaciju bez korištenja ekstremno velikih ljudskih resursa. Uz pomoć kvalitetnih podataka i strojnog učenja, mogli bismo otkriti i nove međuovisnosti koje utječu na akademsku uspješnost, što bi nam omogućilo da u ranoj fazi signaliziramo potencijalne rizike i pružimo personalizirano podršku studentima koji se suočavaju s poteškoćama

Ovaj rad će se baviti uspoređivanjem tehnika strojnog učenja na skupu podataka iz akademske okoline. Uspoređivat će se tri tehnike strojnog učenja, a to su: stabla odlučivanja, metoda k-najbližih susjeda i neuronske mreže. Rad će se baviti teorijskom i praktičnom primjenom navedenih tehnika strojnog učenja. Prilikom treniranja modela optimizirat ću parametre, a zatim ću modele pojedinih metoda koji daju najbolje rezultate međusobno usporediti. Cilj rada je otkriti koje su prednosti i mane pojedinih tehnika strojnog učenja kada se koriste u svrhu generiranja akademskih predikcija. Ukoliko se nađe model koji može generirati akademske predikcije sa dovoljno visokom točnošću, mogao bi biti koristan alat za unaprjeđenje obrazovnog sustava, ali bi mogao i otkriti korelaciju između karakteristika osobe i akademske uspješnosti koju do sada nismo znali.

2. Strojno učenje

U tradicionalnim modelima razvoja softvera, programeri su pisali logiku temeljenu na trenutnom stanju poslovanja te su potom dodavali odgovarajuće podatke. Međutim, promjena u poslovanju postala je norma. Gotovo je nemoguće predvidjeti koje će promjene transformirati tržište. (Hurwitz & Kirsch, 2018). Strojno učenje značajno mijenja dizajn softverskih proizvoda na način da softver čine prilagodljivijim promjenama u poslovanju. Osim što se prilagođava, softver je u mogućnosti izvršavati zadatke bez striktnih uputa korisnika. Poslovna pravila sada sve više proizlaze iz podataka, što omogućuje da se poslovanje bolje nosi sa anomalijama te da se bolje uočavaju obrasci u podatcima. Treba imati u vidu da strojno učenje ima smisla jedino kao dio timskog procesa koji za cilj ima analizu podataka.

2.1. Općenito o strojnom učenju

Strojno učenje nije lako definirati, s obzirom da obuhvaća puno tehnika i algoritama, kao i sam pristup rješavanju problema na način da se omogući stroju da otkrije „*vlastite*“ algoritme bez velike intervencije ljudi. Zajedničko svim definicijama strojnog učenja je postupno poboljšanje sustava na temelju novih opservacija. Jedna valjana definicija strojnog učenja je da je to *grana umjetne inteligencije koja omogućava sustavu da uči iz podataka, umjesto kroz eksplicitno programiranje* (Hurwitz & Kirsch, 2018).

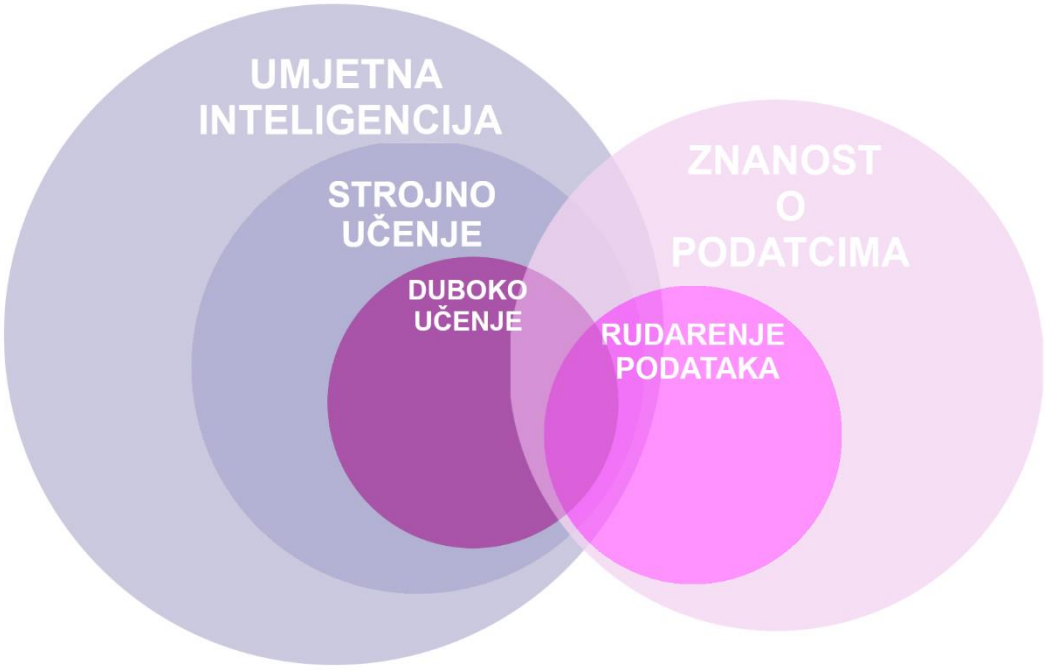
Strojno učenje svoje korijene povlači u statistici, te su neki statistički modeli poput linearne regresije i danas u srži strojnog učenja. *Za pionira strojnog učenja smatramo Alana Turinga, koji je osnove strojnog učenja postavio 50-ih godina prošloga stoljeća, dok strojno učenje poprima današnji oblik 1990-ih godina*. (Wikipedia, 2023)

Strojno učenje koristi veliki broj algoritama kako bi se model poboljšao kroz više iteracija. Tijekom svake iteracije sustav uči kako opisati i analizirati podatke, te predvidjeti buduće ishode. Nakon svake iteracije možemo unijeti u sustav podatke, te će nam na temelju obrazaca naučenih u prethodnoj iteraciji sustav dati rezultat. Na primjer, ulazni podatci mogu biti marka automobila i godina proizvodnje, a rezultat može biti očekivana cijena automobila. Primjer primjene takvih modela su sustavi za preporuku sadržaja, gdje se korisnicima ne prikazuje sadržaj na temelju unaprijed programirane logike, nego na temelju prikupljenih podataka o korisniku.

Kako bi iteracija bila uspješna važno je da podatci budu točni i značajni (Hurwitz & Kirsch, 2018). Također, potrebno je razumjeti izvore podataka, ima li ih smisla kombinirati te koliko ih je potrebno obraditi prije nego pomoću njih krenemo trenirati modele strojnog učenja. Treba

imati na umu da modeli strojnog učenja barataju s brojevima, tako se svi podatci moraju iskazati u brojčanom obliku. U prethodnom primjeru gdje model predviđa cijenu automobila, marku automobila bi morali iskazati kao znamenku ili vektor znamenki, a ne kao riječ.

Postoji puno disciplina koje si često spominju u kontekstu strojnog učenja te je dobro makar približno znati u kojem su međusobnom odnosu, kako bi mogli bolje razumjeti ulogu strojnog učenja u svijetu obrade podataka. Neke od tih disciplina su: umjetna inteligencija, duboko učenje, znanost o podatcima(*engl. data science)* i *rudarenje podatcima(engl. data mining).* Ne postoji univerzalna klasifikacija ovih disciplina, a jedna valjana podjela prikazana je na slici 1.



Slika 1. Vennov dijagram polja srodnih strojnom učenju
(Prema Sveučilište Helsinki & MinnaLearn. (2018). Elements of AI.)

U svom najjednostavnijem obliku, umjetna inteligencija je polje koje kombinira računalne znanosti i robusne skupove podataka u svrhu rješavanja problema. (IBM, 2023) Umjetna inteligencija je široki pojam koji obuhvaća više disciplina među kojima su strojno i duboko učenje, ali i razumjevanje govora, računalni vid i slično. *Najšire gledano, umjetna inteligencija može se shvatiti kao sustavi koji mogu "razmišljati".* (Hurwitz & Kirsch, 2018).

Duboko učenje je podskup strojnog učenja. *Duboko učenje zapravo se sastoji od neuronskih mreža. "Duboko" u dubokom učenju odnosi se na neuronsku mrežu koja se sastoji od više od tri sloja* (IBM, 2023). Duboko učenje je osnova brojnih izuzetno kompleksnih i učinkovitih sustava poput AlphaGo i ChatGPT. *Duboko učenje trenutno je najčešće korišten pristup za stvari poput prepoznavanja vizualnih objekata, strojnog prevođenja, prepoznavanja govora, sinteze govora i stvaranja slika.* (Russell & Norvig, 2021)

Znanost o podacima je noviji pojam koji pokriva nekoliko poddisciplina, a to su strojno učenje, statistika, određene aspekte računalne znanosti uključujući algoritme, pohranu podataka i razvoj web aplikacija. (University of Helsinki & MinnaLearn, bez. dat.) Iako znanost o podatcima koristi strojno učenje i duboko učenje, veliku ulogu imaju i poslovni analitičari te drugi stručnjaci koji razumiju domenu u kojoj se primjenjuju navedene tehnologije.

Rudarenje podatcima je proces istraživanja i analiziranja velikih količina podatka u svrhu uzoraka u tim podacima, a temelji se na načelima statistike. (Hurwitz & Kirsch, 2018). Za razliku od strojnog učenja, rudarenje podataka se ne koristi za predviđanje podataka i potvrđivanje hipoteza, nego za razumjevanje i tumačenje podataka.

Sve navedene discipline temelje se na računalnim znanostima, statistici i obradi podataka, ali se razlikuju u opsegu, svrsi i primjeni. U svrhu ostvarivanja punog potencijala, poželjno je da proces obrade i analize podataka bude dio timskog procesa koji uključuje stručnjake iz svih navedenih disciplina.

2.2. Tipovi strojnog učenja

Postoji više tipova strojnog učenja, a njihov pristup se razlikuje ovisno o tome za koju vrstu problema su namijenjeni. Za razumijevanje podjela potrebno je znati kako se odvija uobičajeni proces treniranja modela strojnog učenja. *Računalo promatra neke podatke, gradi model temeljen na podacima te koristi model kao hipotezu o svijetu, ali i kao dio softvera koji može rješavati probleme* (Russell & Norvig, 2021). Podjele se baziraju na više faktora: prethodno znanje agenta (računala), dostupnost povratne informacije, tip izlaznog podatka i kontinuitet prilagodbe novim podatcima.

Ovisno o prethodnom znanju agenta, postoje agenti s malim prethodnim znanjem koji sve nauče iz skupa podataka, te agenti kod kojih je prisutan transfer znanja, odnosno znanje iz jedne domene se prenosi u drugu. Primjerice, ukoliko neuronsku mrežu koja služi za prepoznavanje cestovnih vozila krenemo trenirati za prepoznavanje bicikala, za očekivati je da će brže dostići visoku razinu točnosti u odnosu na potpuno novu neuronsku mrežu. Razlog tomu je što neuronska mreža koje je već trenirana na nekom relativno sličnom skupu podataka ima „*iskustva*“ sa relevantnim karakteristikama poput guma, cestama i slično

Ovisno o dostupnosti povratne informacije postoji: nadzirano učenje, podržano(engl. reinforcement) učenje i nenadzirano učenje.

U nadziranom učenju agent promatra ulazno-izlazne parove i uči funkciju koja se transformira ulazne podatke u izlazne (Russell & Norvig, 2021). Primjerice, skup podataka za model koji predviđa cijene automobila bi bile marke automobila i godina proizvodnje kao ulazni

podatci te pripadajuća cijena kao izlazni. Nakon treniranja model testiramo na novim podatcima kako bi provjerili koliko dobro model generalizira na podatke koje nije prije vidio.

U nenadziranom učenju agent uči obrasce ulaznih podataka bez ikakve eksplicitne povratne informacije. Najčešći zadatak učenja bez nadzora je grupiranje. (Russell & Norvig, 2021) Nenadzirano učenje najbolje je ukoliko postoji izuzetno veliki skup podataka. Primjerice, često se koristi u detekciji neželjene pošte, s obzirom da postoji veliki skup podataka te širok spektar parametara koji čine razliku između regularne i neželjene pošte.

Podržano učenje bazira se na potkrepljenjima, odnosno nagradama i kaznama na temelju kojih agent uči. Za razliku od drugih vrsta učenja ne postoji skup podataka, nego agent nakon niza akcija dobije povratnu informaciju u obliku nagrade ili kazne, ovisno o rezultatu koji je uslijedio nakon tog niza akcija. Na agentu je zatim da utvrdi koje su akcije zaslužne za rezultat. Najčešće se primjenjuje u robotici i igranju igara. Ukoliko bismo htjeli trenirati model da nauči igrati šah, mogli bismo ga programirati da *„razumije“* pravila, a potom ga upariti da igra protiv drugih igrača ili potencijalno drugih modela. Na kraju igre model bi dobio povratno informaciju o tome da li je pobijedio ili izgubio te bi sukladno tome mijenjao stil igre da poveća broj pobjeda.

Ovisno o tipu izlaznog podatka postoje klasifikacijski i regresijski modeli. Kod regresijskih modela izlazni podatak je broj, poput temperature cijene ili slično. S druge strane, klasifikacijski modeli kao izlazni podatak imaju diskretne vrijednost, odnosno izlazni podatci dolaze iz konačnog skupa podataka. Na primjer, izlazni podatak modela za predikciju ishoda šahovskog meča može biti isključivo pobjeda, poraz ili neriješeno.

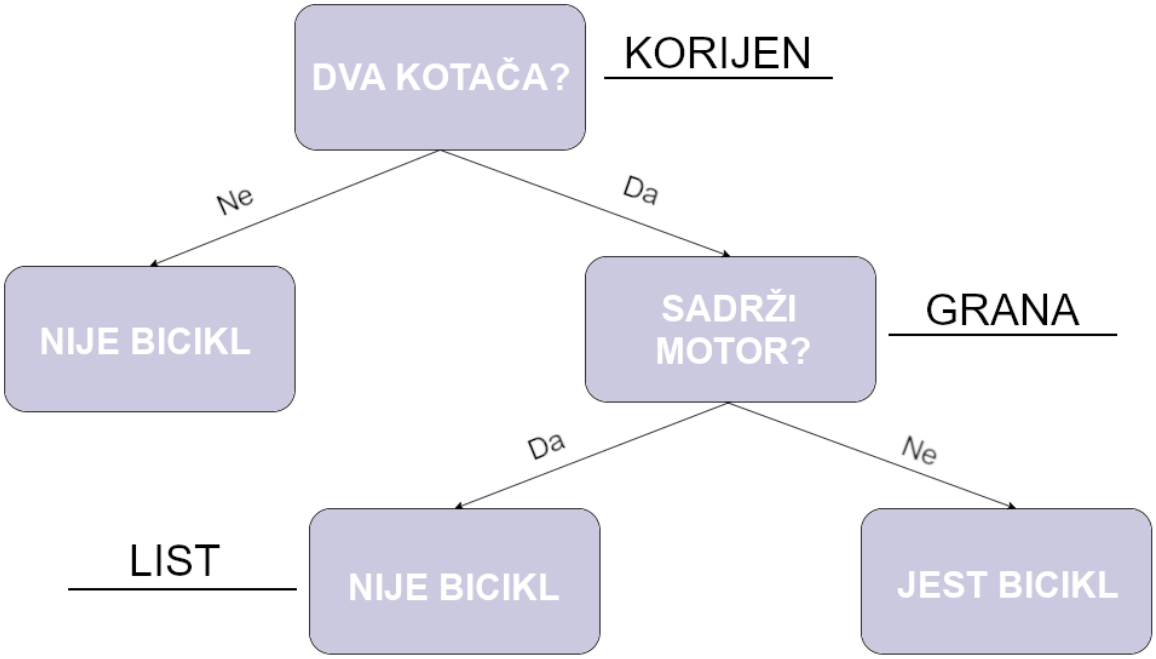
Također, razlikuju se i *online* odnosno *offline* modeli. Razlika je u tome što se online modeli kontinuirano mijenjaju na temelju novih podataka, dok se offline modeli jednom isporučeni više ne mijenjaju.

2.3. Algoritmi strojnog učenja

U ovom radu koristit će se tri tehnike strojnog učenja, a to su stabla odlučivanja, metoda k-najbližih susjeda i neuronske mreže.

2.3.1. Stabla odlučivanja

Stablo odlučivanja je reprezentacija funkcije koja preslikava vektor vrijednosti atributa u jednu izlaznu vrijednost – „odluku“. (Russell & Norvig, 2021) Stablo odlučivanja ima strukturu sličnu uređenim stablima koja se koriste kao strukture podataka u programiranju. Na slici je prikazana struktura stabla odlučivanja koje prepoznaje da li je prijevozno sredstvo bicikl ili nije.



Slika 2 Stablo odluke za prepoznavanje bicikla
(Vlastita izrada)

Stablo odlučivanja donosi odluku, odnosno izlaznu vrijednost tako što obavlja niz testova počevši od korijena preko grana i sve do lista. Svaki čvor (korijen i grana) test je jedne ulazne vrijednost, a svaki list označava odluku, odnosno izlaznu vrijednost. Ulazne vrijednosti mogu biti i kontinuirane varijable, a tada su testovi u obliku raspona. Primjerice u prethodnom modelu bismo mogli dodati čvor koji testira ima li prijevozno sredstvo više od 44,5 kg te ukoliko ima, kao rezultat odlučiti da prijevozno sredstvo nije bicikl.

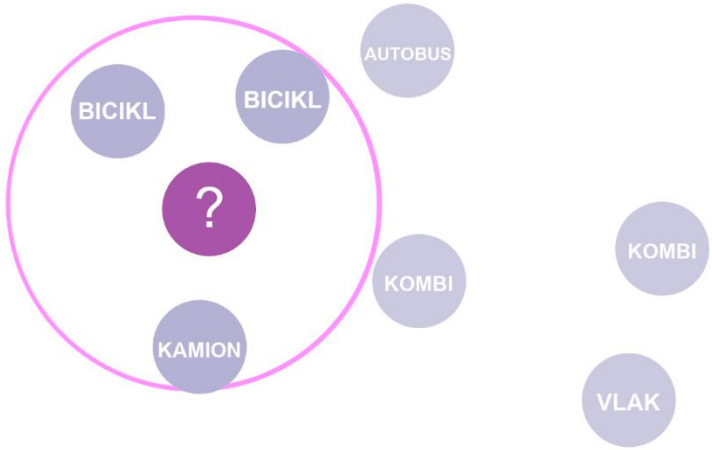
Poželjno je pronaći najmanje moguće stablo odlučivanja koje je i dalje konzistentno s podatcima za treniranje. Iako nije moguće garantirati da ćemo pronaći najmanje moguće stablo, postoje algoritmi koji omogućavaju optimizaciju stabla odlučivanja. *Algoritam izgradnje*

stabla odlučivanja usvaja pohlepnu strategiju podjeli pa vladaj: uvijek prvo testira najvažniji atribut, a zatim rekurzivno rješava manje potprobleme koji su definirani mogućim rezultatima testa. (Russell & Norvig, 2021) Svrha testiranja najvažnijeg atributa na početku je smanjivanje prosječnog broja koraka potrebnih za donošenja odluke. Idealan atribut je onaj koji može klasificirati sve primjere. Također važno je testirati samo one atribute koji donose neku značajnu informaciju. Na primjer, u prethodnom primjeru modela za prepoznavanje bicikala nebi imalo smisla testirati ima li prijevozno sredstvo svjetla, jer bismo opet morali testirati ima li prijevozno sredstvo dva točka te sadrži li motor, stoga je moguće klasificirati sve primjere i bez tog atributa.

Stabla odlučivanja imaju brojne prednosti kao: laku razumljivost, mogućnost prilagodbe velikim skupovima podataka i sposobnost rukovanja diskretnim i kontinuiranim ulazima, kao i mogućnost klasifikacije i regresije. Također, imaju i neke mane poput: suboptimalne točnosti (većinom zbog pohlepnih algoritama), potencijalno dugo vrijeme izvršavanja potrebno za dobivanje izlazne vrijednosti, te nestabilnost prilikom dodavanja novih podataka (jedan podatak može promijeniti strukturu stabla iz korijena)

2.3.2. Metoda k-najbližih susjeda

Metoda k-najbližih susjeda je ne-parametarski model koji donosi klasifikacije ili predikcije na temelju najsličnijih(najbližih) primjera u skupu podataka. Kažemo da je metoda k-najbližih susjeda ne-parametarski jer ne pokušava prikazati skup podataka u obliku funkcije, nego rezultat vraća na temelju najbližih susjeda u skupu podataka. Za klasifikacijske probleme model će jednostavno vratiti oznaku(*engl. label)* najbližih susjeda, dok za regresijske probleme je moguće vratiti prosjek(ili neku drugu funkciju) najbližih susjeda. Na slici 3 prikazana je model k-najbližih susjeda koji prepoznaje da li je prijevozno sredstvo bicikl.



Slika 3 Model k-najbližih susjeda za prepoznavanje bicikla

(Vlastita izrada)

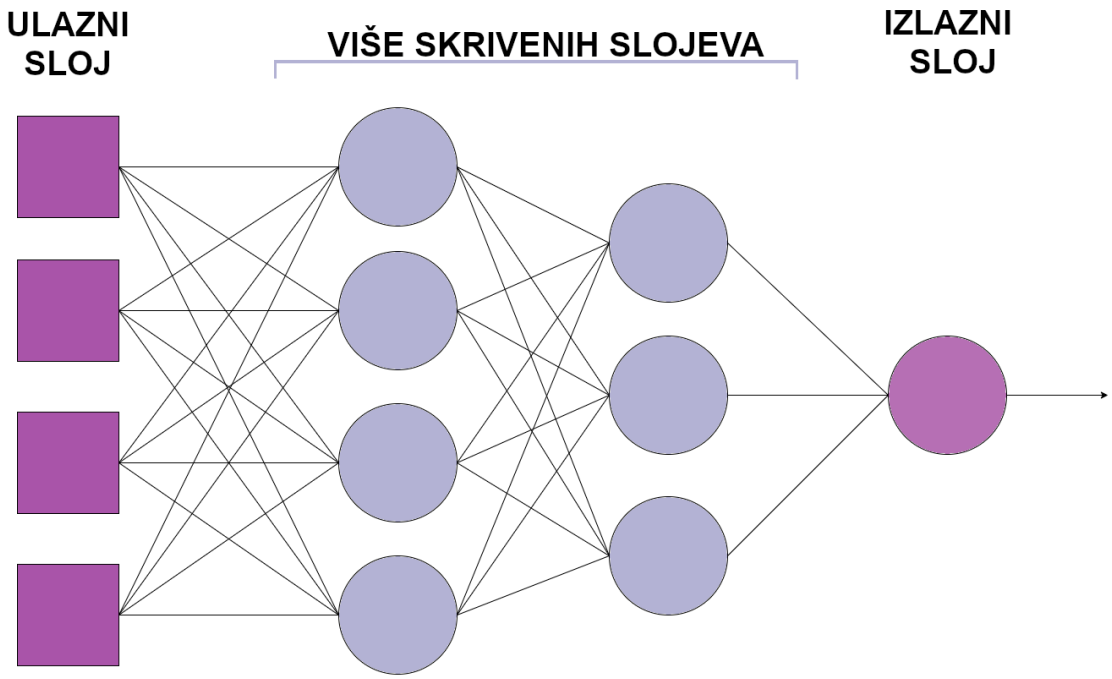
U ovom primjeru u obzir se uzimaju tri najbliža susjeda te bi ovaj model kao rezultat vratio da je prijevozno sredstvo bicikl, s obzirom da su većina najbližih susjeda bicikla. Treba imati na umu da određivanje najbližeg susjeda nije uvijek toliko jednostavno kao računanje geometrijske udaljenosti na primjer, neke ulazne vrijednosti mogu biti tekstualne ili općenito nenumeričke vrijednosti. S druge strane, numeričke vrijednosti bi trebale biti normalizirane. jer u suprotnom bi iste veličine iskazane u drugim mjernim jedinicma mogle značajno drugačije utjecati na rezultat. Primjerice, atribut iskazan u kilometrima bi značajno manje utjecao na računanje najbližeg susjeda u odnosu na isti atribut iskazan u centimetrima. *Postoji i* više metrika za mjerenje udaljenosti poput: Euklidove udaljenosti, Manhattan udaljenosti, Minkowski udaljenosti i drugih, a odabir tehnike računanja udaljenosti ovisi od slučaja do slučaja.

Nadalje, ne želimo uvijek svim susjedima koji su blizu našoj točki davati jednaku važnost. Željeli bismo da najbliži susjedi imaju najveći utjecaj na izlaznu vrijednost, oni dalji nešto manji utjecaj, dok najdalji uopće nemaju utjecaj na izlaznu vrijednost. (Russell & Norvig, 2021) Ovako bismo model učinili više kontinuiranim, pogotovo u rubnim slučajevima gdje je manje bliskih susjeda. Važnost svakog susjeda određujemo pomoću funkcije zvana jezgra (engl. kernel) čiji je argument udaljenost između naš točke i susjeda. (Russell & Norvig, 2021)

Prednosti metode k-najbližih susjeda su jednostavnost implementacije, pogotovu u usporedbi s drugim algoritmima i dobra prilagodba novim podatcima jer je dovoljno samo nove primjere dodati u skup podataka za treniranje. S druge strane, metoda postaje izuzetno zahtjevna(u smislu vremena izvođenja i memorije) sa povećanjem skupa podataka, loše se nosi sa velikim brojem atributa te je sklona loše generalizirati na nove podatke. (IBM, 2023)

2.3.3.Neuronske mreže

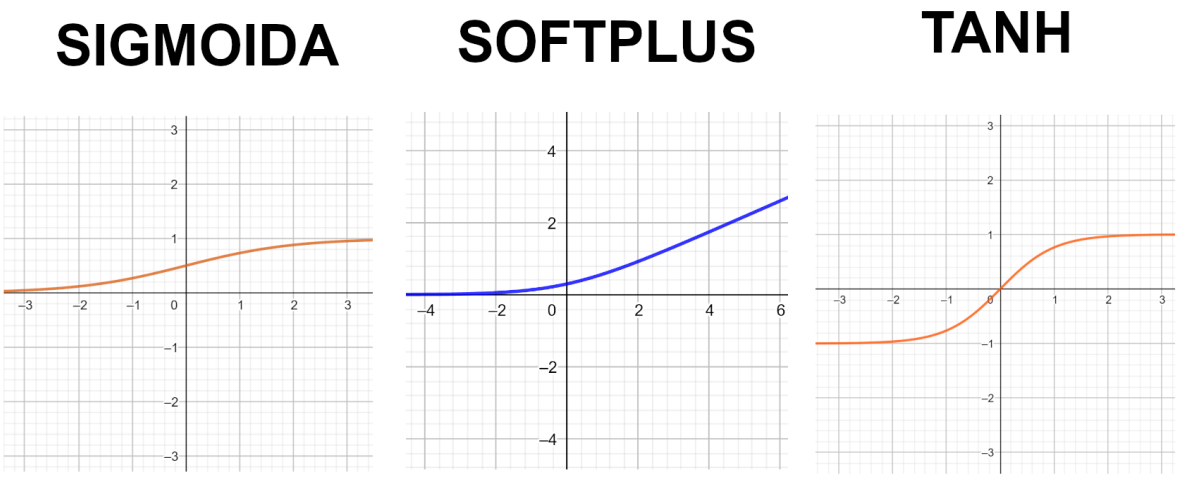
Struktura i naziv neuronskih mreža inspirirana je neuronima u mozgu koji komuniciraju međusobno, s time da je sličnost sa pravim neuronima na apstraktnom nivou. Neuronske mreže sadrže ulazni sloj, jedan ili više skriven te izlazni sloj. Slojevi su sačinjeni od „neurona“. Svaki čvor(neuron) ima svoju težinu, prag te aktivacijsku funkciju. Prilikom računanja, sve ulazne vrijednosti čvora se pomnože pripadajućim težinama te se njihov zbroj koristi kao argument aktivacijske funkcije. Ukoliko rezultat aktivacijske funkcije pređe prag neurona, rezultat aktivacijske funkcije pomnožen pripadajućom težinom bit će korišten kao ulaz u drugi neuron. Ovaj proces se ponavlja sve do izlaznog sloja, gdje dobijemo rezultate neuronske mreže. Struktura neuronske mreže prikazan je na slici 4.



Slika 4 Struktura neuronske mreže

(Prema Russell & Norvig, 2021)

Aktivacijske funkcije u čvorovima su nelinearne, jer u protivnom kompozicija čvorova bi i dalje bila linearna funkcija. Zbog toga, model može gotovo savršeno reprezentirati bilo koji skup podataka Neke od najčešćih aktivacijskih funkcija su sigmoida, softplus i tahn funkcija, čiji su grafovi na slici 5.



Slika 5. Grafovi funkcija

(Vlastita izrada)

Neuronske mreže uče algoritmom širenja unatrag(*engl. backpropagation).* Ovaj algoritam prvo izračuna rezultat neuronske mreže zatim, ovisno o grešci, konfigurira težine neurona sve dok se ne postigne željena točnost. Ovaj algoritam važan je za odabir aktivacijskih funkcija. *Do 2010. najviše su se koristile sigmoida i tanh funkcija, no s obzirom da algoritam širenja unatrag bolje funkcionira s softplus i ReLu funkcijom, one su sada najzastupljenije aktivacijske funkcije.* (Russell & Norvig, 2021)

Neuronske mreže najčešće su korištene za probleme gdje je postoji veliki skup podataka, kao i puno parametara(i do nekoliko milijardi). Primjerice, često se koriste za računalni vid, procesiranje govora, generiranje slika i slično.

3. Opis problema

U prethodnim poglavljima, objašnjene su teorijske osnove strojnog učenja i tehnika koje ćemo koristiti, no u praktičnoj primjeni strojnog učenja potrebno je uzeti u obzir dodatne okolnosti. Ne postoji univerzalna metodologija za razvijanje sustava strojnog učenja, no neki od uobičajenih procesa koje je potrebno provesti su: formulacija problema, prikupljanje i obrada podataka, analiza podataka, treniranje i testiranje modela.

3.1. Formulacija problema

Za uspješno formuliranje problema potrebno je odgovoriti na dva pitanja: „koji problem želimo riješiti potencijalnom korisniku“ i „koji dio tog problema možemo riješiti pomoću strojnog učenja“. (Russell & Norvig, 2021). Potencijalnom korisniku željeli bismo omogućiti generiranje predikcije akademske uspješnosti koja bi pomogla u procjeni akademskog potencijala. Strojnim učenjem možemo konstruirati model koji će generirati pretpostavku o uspješnosti završetka studija na temelju podataka o studentu.

Za detaljniju specifikaciju potrebno je odlučiti koji tip podataka će biti predikcija koju model generira. Intuitivno bi bilo da model predviđa prosjek ocjena studenta, s obzirom da ocjene predstavljaju koliko je uspješno student položio kolegije. Ipak, problem se javlja ukoliko je u skupu podataka više fakulteta od kojih svaki fakultet ima vlastite kriterije ocjenjivanja. Također, postoji problem sa rangiranjem uspješnosti studenta koji ima bolji prosjek ocjena, ali ima manje položenih predmeta u odnosnu na drugog studenta. Univerzalnija metrika mjerenja uspješnosti studenta je vrijeme studiranja studenta u odnosu na propisano trajanje studija. U ovom radu predikcija će biti u obliku kategorijske oznake (*engl.* label) sa tri moguće vrijednosti: završio studij, upisan u studij te odustao od studija. Detaljnija objašnjenja vidljiva su u tablici 1.

Dio formuliranja problema je i kategorizirati problem strojnog učenja(Poglavlje 2.2). S obzirom da će predikcija modela biti u obliku kategorijske oznake ovo spada u problem klasifikacije. Također, skup podataka koje će biti korišten za treniranje modela sadrži ulazne atribute kao i kategorijsku oznaku studenta, stoga ovo spada u problem nadziranog učenja. Model također neće imati nikakvo prethodno znanje te se neće prilagođavati novim podatcima

Oznaka	Značenje
Završio	Student je nakon propisanog trajanja studija uspješno završio studijski program
Upisan	Student je nakon propisanog trajanja studija i dalje upisan u studij, odnosno student studira dulje od propisanog trajanja studija
Odustao	Student je odustao od studijskog programa koji je upisao. Studenti koji su promijenili studij također spadaju u ovu kategoriju

Tablica 1 Kategorijske oznake predikcije modela

(Vlastita izrada)

3.2. Prikupljanje i obrada podataka

U ovom radu će biti korišteni podatci preuzeti s web stranice Kaggle. Kaggle je online platforma koja služi za dijeljenje skupova podataka, održavanje natjecanja u strojnom učenju i sličnim disciplinama, a danas je jedna od najvećih online zajednica podatkovnih znanstvenika. Sami podatci su rezultat rada *Predicting Student Dropout and Academic Success* portugalskih znanstvenika Valentim Realinha, Jorge Machada, Luís Baptiste i Mónice Martins objavljenog 2022. godine. U tom radu kao izvor podataka navedene su portugalske ustanove koje bilježe demografske, socioekonomske, makroekonomske i akademske podatke. U te ustanove spadaju: Politehnički institut Portalegre, Generalna uprava za visoko obrazovanje(DGES), ustanova za Nacionalno natjecanje za pristup visokom obrazovanju(CNAES), Suvremena portugalska baza podataka(PORDATA).

Podaci se odnose na portugalske studente koji su upisali studij između 2008. i 2019. godine te sadrži podatke sa 17 različitih preddiplomskih studija, poput agronomije, dizajna, obrazovanja, sestrinstva, novinarstva, menadžmenta, socijalne skrbi i tehnologije. Konačni skup podataka dostupan je kao CSV datoteka s 4424 zapisa, 35 atributa i ne sadrži nepopunjene vrijednosti.

Proces prikuplja i obrade podataka uključivao je više koraka. Prvi korak bila je transformacija Microsoft Access baze podataka u CSV datoteku, koja je sadržavala podatke o

studentima prilikom upisa na fakultet. Drugi korak bila je ekstrakcija demografskih i socioekonomskih atributa studenata, što je uključivalo brisanje duplikata i nerelevantnih studenata te brisanje i preimenovanje nekih atributa. Treći korak bio je dohvaćanje podataka vezanih uz prve dvije godine preddiplomskog studija. Konačno, svi podatci se sjedine te pretprocesiraju. Pretprocesiranje podataka za cilj je imalo razriješiti anomalije, stršila te vrijednosti koje nedostaju. (Realinho, Machado, Baptista, & Martins, 2022) Na kraju je svaki student dobio oznaku sukladno tablici 1.

3.3. Opis skupa podataka

Skup podataka sadrži kontinuirane i diskretne vrijednosti. Tekstualni vrijednosti transformirane su u numeričke vrijednosti. Tablica 2 sadrži: ime i moguće vrijednosti atributa.

Ime atributa	**Moguće vrijednosti**
Bračni status	1—Samac 2—Oženjen/Oženjena 3—Udovac/Udovica 4—Razveden/Razvedena 5—Izvanbračna zajednica 6—Pravno razdvojen/Pravno razdvojena
Tip prijave	1—1. faza—opći kontingent 2—Uredba br. 612/93 3—1. faza—poseban kontingent (Otok Azori) 4—Diplomanti drugih viših tečajeva 5—Uredba br. 854-B/99 6—Međunarodni student (preddiplomski) 7—1. faza—poseban kontingent (Otok Madeira) 8—2. faza—opći kontingent 9—3. faza—opći kontingent 10—Uredba br. 533-A/99, točka b2) (Različiti Plan) 11—Uredba br. 533-A/99, točka b3 (Druga Institucija) 12—Stariji od 23 godine 13—Prijenos 14—Izmjena smjera/studija 15—Diplomanti tehnološke specijalizacije 16—Izmjena institucije/smjera 17—Diplomanti kratkog ciklusa 18—Izmjena institucije/smjera (Međunarodno)
Broj prijave (redni)	Cijeli pozitivni broj
Fakultet	1—Tehnologije proizvodnje biogoriva 2—Animacija i dizajn multimedije 3—Socijalna služba (večernja nastava) 4—Agronomija 5—Dizajn komunikacija 6—Veterinarska njega 7—Informatički inženjering 8—Konjičarstvo 9—Menadžment

	10—Socijalna služba 11—Turizam 12—Medicinska sestra 13—Oralna higijena 14—Upravljanje oglašavanjem i marketingom 15—Novinarstvo i komunikacija 16—Osnovno obrazovanje 17—Upravljanje (večernja nastava)
Dnevna/noćna nastava	1—dnevno 0—večernje
Prethodno obrazovanje	1—Srednje obrazovanje 2—Više obrazovanje—preddiplomski stupanj 3—Više obrazovanje—diploma 4—Više obrazovanje—magisterij 5—Više obrazovanje—doktorat 6—Različite razine višeg obrazovanja 7—12. godina školovanja—nedovršeno 8—11. godina školovanja—nedovršeno
Nacionalnost	1—Portugalska 2—Njemačka 3—Španjolska 4—Talijanska 5—Nizozemska 6—Engleska 7—Litvanska 8—Angolska 9—Zelenortska (Cape Verdean) 10—Gvinejska 11—Mozambička 12—Santomski 13—Turska 14—Brazilka 15—Rumunjska 16—Moldavija 17—Meksička 18—Ukrajinska 19—Ruska 20—Kubanska 21—Kolumbijska
Majčino obrazovanje	1—Srednje obrazovanje—12. godina školovanja ili ekvivalent 2—Više obrazovanje—preddiplomski stupanj 3—Više obrazovanje—diploma 4—Više obrazovanje—magisterij 5—Više obrazovanje—doktorat 6—Učestalost višeg obrazovanja 7—12. godina školovanja—nedovršeno 8—11. godina školovanja—nedovršeno 9—7. razred (stari sustav) 10—Drugo—11. godina školovanja 11—2. godina komplementarnog srednjoškolskog tečaja 12—10. godina školovanja 13—Opći trgovački tečaj

	14—Osnovno obrazovanje 3. ciklus (9./10./11. razred) ili ekvivalent 15—Komplementarni srednjoškolski tečaj 16—Tehničko-zanatski tečaj 17—Komplementarni srednjoškolski tečaj—nedovršeno 18—7. razred školovanja 19—2. ciklus općeg srednjoškolskog tečaja 20—9. godina školovanja—nedovršeno 21—8. godina školovanja 22—Opći tečaj administracije i trgovine 23—Dopunska računovodstvena i administracija 24—Nepoznato 25—Ne može čitati ni pisati 26—Može čitati bez završene 4. godine školovanja 27—Osnovno obrazovanje 1. ciklus (4./5. godina) ili ekvivalent 28—Osnovno obrazovanje 2. ciklus (6./7./8. godina) ili ekvivalent 29—Tečaj tehnološke specijalizacije 30—Više obrazovanje—diploma (1. ciklus) 31—Specijalizirani visokoškolski studijski tečaj 32—Profesionalni visokoškolski tehnički tečaj 33—Više obrazovanje—magisterij (2. ciklus) 34—Više obrazovanje—doktorat (3. ciklus)
Očevo obrazovanje	Isto kao i majčino obrazovanje
Majčino zanimanje	1—Student 2—Predstavnici zakonodavne vlasti i izvršnih tijela, direktori, rukovoditelji i izvršni menadžeri 3—Specijalisti u intelektualnim i znanstvenim djelatnostima 4—Tehničari i zanimanja srednjeg nivoa 5—Administrativno osoblje 6—Osoblje za osobne usluge, sigurnost i zaštitu te prodavači 7—Poljoprivrednici i vješti radnici u poljoprivredi, ribarstvu i šumarstvu 8—Vješti radnici u industriji, građevini i obrtnicima 9—Radnici za instalacije, operateri strojeva i montažeri 10—Nekvalificirani radnici 11—Zanimanja u oružanim snagama 12—Druge situacije; 13—(prazno) 14—Časnici oružanih snaga 15—Narednici oružanih snaga 16—Ostalo osoblje oružanih snaga 17—Ravnatelji administrativnih i trgovačkih usluga 18—Ravnatelji hotela, ugostiteljstva, trgovine i drugih usluga 19—Specijalisti u fizikalnim znanostima, matematici, inženjeringu i srodnim tehnikama 20—Zdravstveni stručnjaci 21—Učitelji 22—Specijalisti za financije, računovodstvo, administrativnu organizaciju i javne i komercijalne odnose 23—Tehničari srednjeg nivoa u znanosti i inženjeringu 24—Tehničari i profesionalci srednjeg nivoa u zdravstvu 25—Tehničari srednjeg nivoa iz pravnih, društvenih, sportskih, kulturnih i sličnih usluga

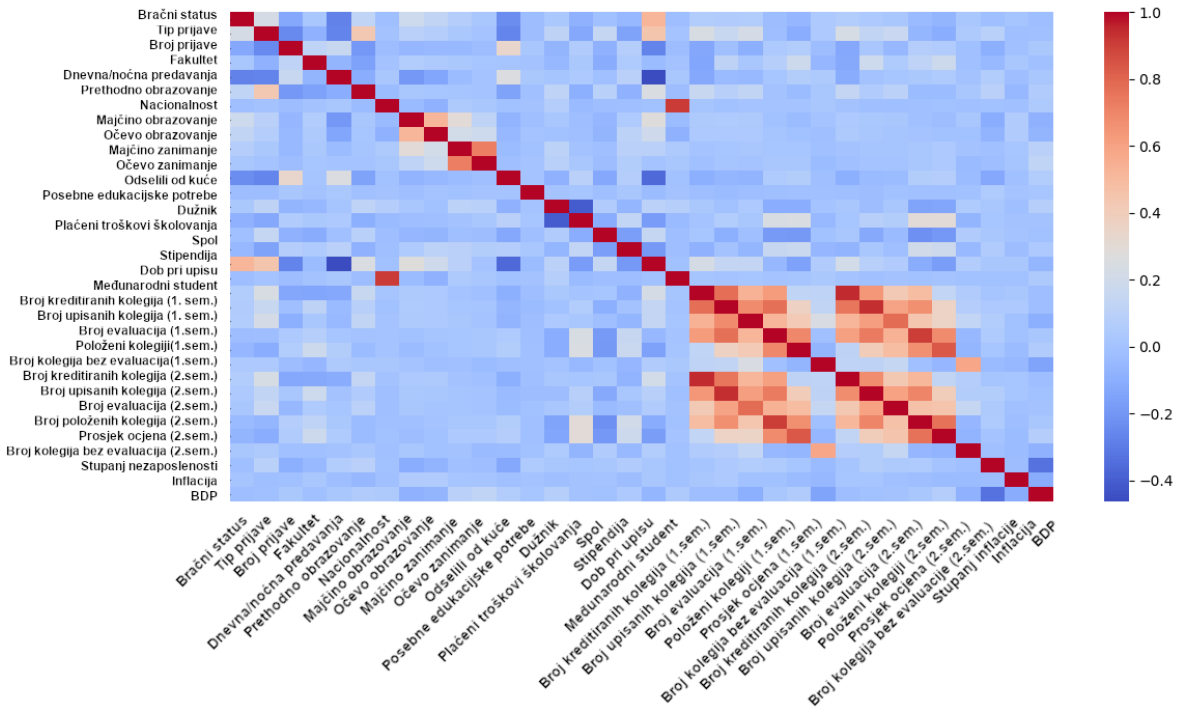
	26—Tehničari za informacijske i komunikacijske tehnologije 27—Uredski radnici, tajnici općenito i operateri obrade podataka 28—Operateri za obradu podataka, računovodstveni, statistički, financijski usluga i usluga vezanih uz registre 29—Ostalo administrativno osoblje za podršku 30—Radnici osobnih usluga 31—Prodavači 32—Radnici za osobnu njegu i slično 33—Osoblje za zaštitu i sigurnost 34—Poljoprivrednici usmjereni na tržište i vješti radnici u poljoprivredi i stočarstvu 35—Poljoprivrednici, stočari, ribari, lovci i sakupljači, te subsistencija 36—Vješti radnici u građevini i slično, osim električara 37—Vješti radnici u metalurgiji, obradi metala i slično 38—Vješti radnici u elektrotehnici i elektronici 39—Radnici u preradi hrane, drvnoj industriji, odjeći i drugim industrijama i zanatima 40—Radnici na stacionarnim postrojenjima i strojevima 41—Montažeri 42—Vozači vozila i operatori pokretnih uređaja 43—Nekvalificirani radnici u poljoprivredi, proizvodnji životinja i ribarstvu, te šumarstvu 44—Nekvalificirani radnici u ekstraktivnoj industriji, građevini, proizvodnji i transportu 45—Pomoćnici za pripremu obroka 46—Prodavači na ulici (osim hrane) i pružatelji usluga na ulici
Očevo zanimanje	Isto kao majčino zanimanje
Odselio od kuće	1—Da 0—Ne
Posebne edukacijske potrebe	1—Da 0—Ne
Dužnik	1—Da 0—Ne
Plaćeni troškovi školovanja	1—Da 0—Ne
Spol	1—Žena 0—Muškarac
Stipendija	1—Da 0—Ne
Dob pri upisu	Cijeli nenegativni broj
Međunarodni student	1—Da 2—Ne
Broj kreditiranih kolegija (1. semestar)	Cijeli nenegativni broj
Broj upisanih kolegija (1.semestar)	Cijeli nenegativan broj
Broj evaluacija (1.semestar)	Cijeli nenegativan broj
Broj položenih kolegija(1.semestar)	Cijeli nenegativan broj
Prosjek ocjena (1.semestar)	Decimalni broj od 1 do 20

Kolegiji bez evaluacija (1. semestar)	Cijeli nenegativni broj
Broj kreditiranih kolegija (2. semestar)	Cijeli nenegativni broj
Broj upisanih kolegija (2.semestar)	Cijeli nenegativan broj
Broj evaluacija (2.semestar)	Cijeli nenegativan broj
Broj položenih kolegija(2.semestar)	Cijeli nenegativan broj
Prosjek ocjena (2.semestar)	Decimalni broj od 1 do 20
Kolegiji bez evaluacija (2. semestar)	Cijeli nenegativni broj
Stupanj nezaposlenosti	Decimalni nenegativni broj
Inflacija	Decimalni broj (postotak)
BDP	Decimalni broj (postotak)

Tablica 2. Naziv i vrijednosti atributa

(Prema Realinho, Machado, Baptista, & Martins, 2022)

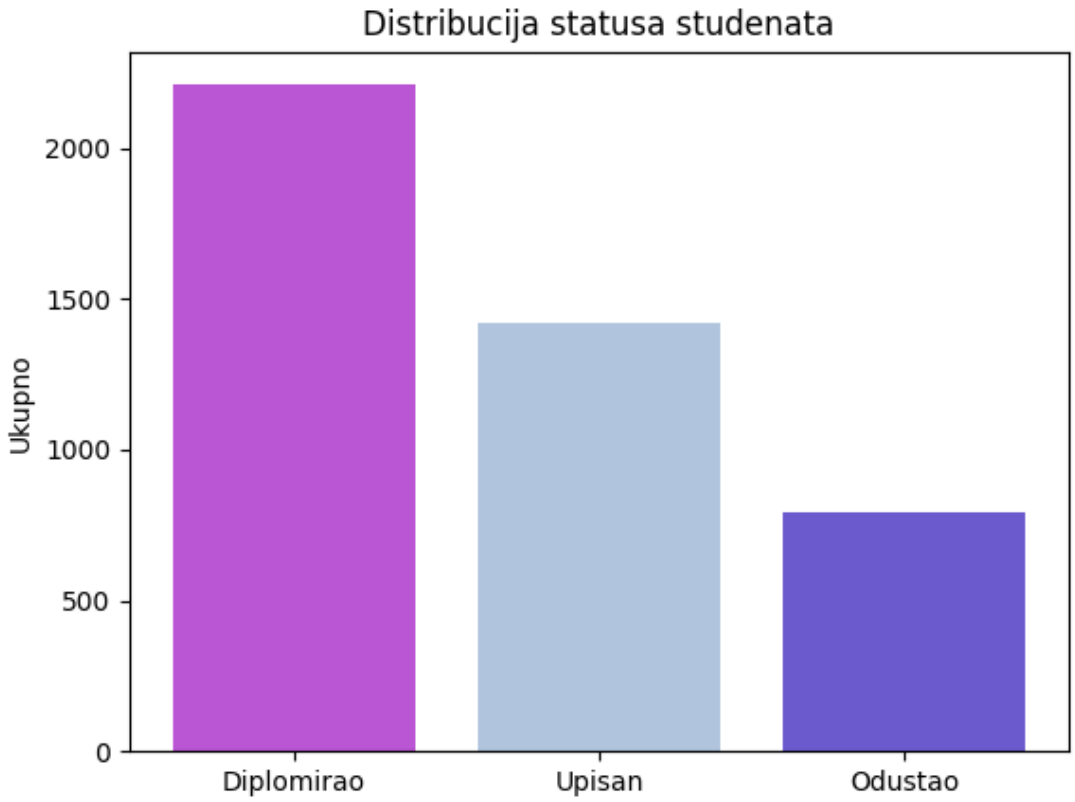
Za bolje razumijevanje atributa i korelacija korisno je izračunati Pearsonov koeficijent. To je koeficijent koji govori koliko su dva atributa međusobno ovisna, vrijednosti bliže 1 označavaju pozitivnu linearnu ovisnost, vrijednost bliže -1 označavaju negativnu linearnu ovisnost, dok vrijednosti 0 označavaju ne postajanje linearne ovisnosti. Na temelju izračunatih vrijednosti konstruirana je toplinska mapa (*engl. heatmap)* gdje toplije boje sugeriraju veću korelaciju.



Slika 6 Toplinska mapa atributa

(Prema Realinho, Machado, Baptista, & Martins, 2022)

Na slici 6 vidljivo je da su najveće korelacije između atributa vezanih za akademsku performansu tijekom prva dva semestra te između majčinog i očevog zanimanja. Kao što je prethodno navedeno, svi studenti su kategorizirani sukladno tablici 1, a distribucija je prikazana na slici 7.



Slika 7 Distribucija prema statusu studenta

(Vlastita izrada)

3.4. Treniranje i testiranje modela

Treniranje i odabir najboljih modela svodi se na jednostavan algoritam. Inicijalni model svake od tehnika strojnog učenja bit će jednostavna inačica modela, primjerice stablo odlučivanja koje sadrži samo korijen. Potom će se postupno trenirati kompleksnije verzije, prilikom čega će se voditi računa s kolikom točnošću model generira predikcije. Kada se razina točnosti modela prestane značajno poboljšavati ili se krene pogoršavati, najbolji modeli svake od tehnika bit će međusobno uspoređeni.

Prilikom treniranja i testiranja modela potrebno je podijeliti skup podatke na više manjih skupova. Ukoliko treniramo samo jedan model, dovoljna su dva skupa podataka, jedan za treniranje, a drugi za testiranje podataka. Ukoliko bi isti skup bio korišten za treniranje i

testiranje podataka, model bi potencijalno mogao parametre pretjerano prilagoditi skupu podataka za treniranje, što bi vjerojatno značilo da loše generalizira na nove podatke. *Kod treniranja više različitih modela poželjno je podijeliti skup podataka na tri skupa: skup za treniranje, validaciju i testiranje.* (Russell & Norvig, 2021). Skup za validaciju bit će korišten za optimizaciju pojedinih modela tehnika strojnog učenja, dok će skup za testiranje biti korišten za usporedbu različitih modela tehnika strojnog učenja.

Prilikom validacije i testiranja modela metrika točnosti bit će F1 rezultat. F1 rezultat sadrži dvije komponente, preciznost i opoziv (Korstanje, 2021) . Formula za preciznost je:

$$\frac{\#\ Pravi\ pozitivni}{\#\ Pravi\ pozitivni + \#\ Lažni\ pozitivni}$$

Preciznost je u biti pokazatelj koji govori koji je omjer nađenih pravih pozitivnih slučajeva u odnosu na ukupan pretpostavljeni broj pozitivnih slučajeva. Opoziv je omjer nađenih pozitivnih slučajeva u odnosu na stvarni broj pozitivnih slučajeva. Formula je:

$$\frac{\#\ Pravi\ pozitivni}{\#Pravi\ pozitivi + Lažni\ negativni}$$

Konačno formula za F1 rezultat je:

$$2 \times \frac{Preciznost \times Opoziv}{Preciznost + Opoziv}$$

3.5. Prethodne studije

Postoji velik broj studija koje se bave analizom podataka kako bi se dobio uvid u faktore koji su dobri indikatori akademske performanse.

Problemom predikcije akademske uspješnosti student bavi se rad *Using learning analytics to develop early warning system for at-risk students* (Akçapınar, Altun, & Aşkar, 2019). Cilj rada bio je predvidjeti uspješnost studenata na kolegiju Računalni hardver (*engl. Computer Hardver)* koristeći više tehnika. Također, cilj je bio usporediti točnost predikcija generiranih na temelju podataka na početku i na kraju semestra. Najveću točnost imao je model tehnike K-najbližih susjeda koji je predviđao ocjene studenata na kolegiju sa točnošću od 83%, dok su neuronske mreže imale točnost od 82%, ostale metode oko 80%. Predikcije temeljene na podatcima s početka semestra imale su manju točnost za približno 10%.

Sličan ovome je rad *Prediction of student academic performance using Moodle data from a Further Education setting* sa Tehnološkog sveučilišta Dublin. Tom radu je također bio cilj analizirati performanse modela ovisno u dostupnim podatcima. Najbolji metoda je bila metoda nasumičnih šuma, a za kvalitetnu predikciju bili su potrebni podatci s Moodlea od minimalno prvih 6 tjedana.

Predikcijom akademske uspješnosti pomoću strojnog učenja bavi se i rad *Predicting Student Performance using Advanced Learning Analytics* predstavljen 2017 na Meuđunarodnoj World Wide Web konferenciji. Ovaj rad u fokus stavlja obiteljsku imovinu i troškove studenta kao skup podataka korišten za predikciju. Kao najuspješnija metoda ispostavila se metoda potpornih vektora(SVM).

U radu Predikcija uspješnosti studenata primjenom umjetnih neuronskih mreža od strane Tea Ljubičića i Dr. sc. Marka Hella ispitane su mogućnosti umjetnih neuronskih mreža u predikciji akademske uspješnosti. Kao metrika za mjerenje uspješnosti studiranja uzet je konačni prosjek na fakultetu. Konačni model je generirao predikcije sa standardnom devijacijom od 0.2.

Juan L. Rastrollo-Guerrero, Juan A. Gómez-Pulido i Arturo Durán-Domínguez u radu Analyzing and Predicting Students' Performanceby Means of Machine Learning: A Review analiziraju 70 radova na temu predikcije akademske uspješnosti. U radu je istaknuto kako se većina radova bazira na sveučilišnom uspjehu, dok za ostala područja nema previše istraživanja. Najčešće i najbolje metode korištene za predikciju su bile metoda potpornih vektora, stabla odlučivanja i nasumične šume.

Postoji više završni i diplomskih radova na Fakultetu organizacije i informatike koji se bave predikcijom i poboljšanjem akademske uspješnosti. Jedan takav je *Ispitivanje uspješnosti studiranja primjenom tehnika rudarenja podataka* (Jurman, 2016). Kao najvažniji faktori za akademsku uspješnost navedeni su: broj utrošenih sati na učenje, uloženi trud, zainteresiranost za fakultet i materijali iz kojih student uči. Najučinkovitija metoda za predikciju bile su neuronske mreže

4. Ispitivanje uspješnosti studiranja

U nastavku će biti implementirane prethodno navedene tehnike strojnog učenja. Za implementaciju će se koristiti Python 3.11.2, pretežito biblioteke Pandas 2.0.3, numpy 1.25.2 i scikit-learn 1.3.0.

4.1. Stabla odlučivanja

Implementacija stabla odlučivanja je prilično jednostavna koristeći navedene biblioteke. Prvo je potrebno uvesti sve biblioteke koje ćemo koristiti u modeliranju.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix
```

Potom čitamo podatke iz excel tablice koja sadrži naš skup podataka, a zatim enkodiramo stupac koji sadrži status studenta. Potom dijelimo skup podataka na način na koji je opisano u prethodnom poglavlju.

```
DATAPATH = r"C:\Users\Nikola\Downloads\archive\dataset.csv"

data = pd.read_csv(DATAPATH)

le = LabelEncoder()


y = le.fit_transform(data['Status'])
X = data.drop('Status', axis=1)

X_train_temp, X_test, y_train_temp, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
X_train, X_val, y_train, y_val = train_test_split(X_train_temp,
y_train_temp, test_size=0.25, random_state=42)
```

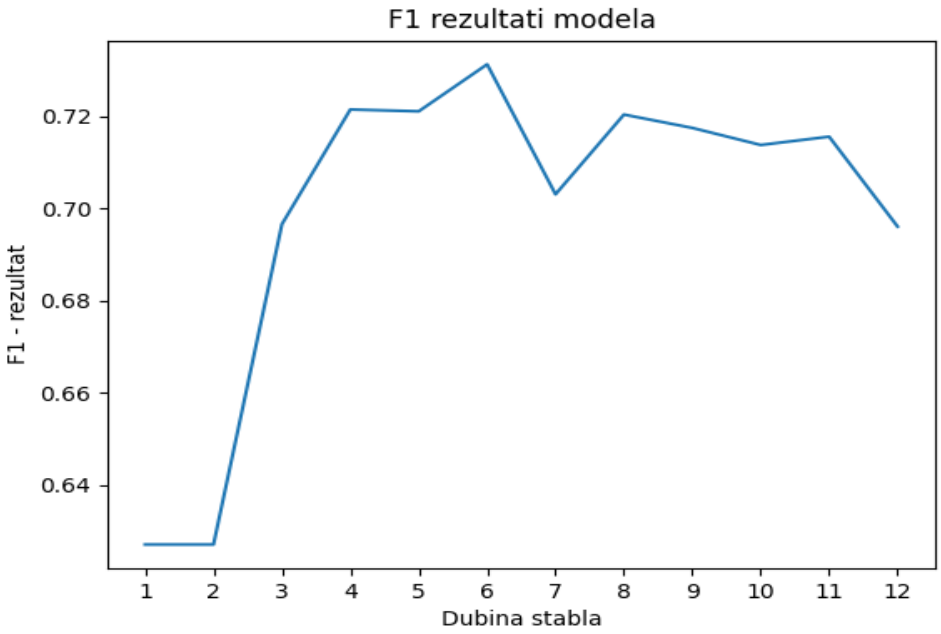
Slijedi samo testiranje modela. Atribut max_depth je takozvani hiperparametar, odnosno parametar koji konfiguriramo prije treniranja modela te pripada cijeloj klasi, a ne samom modelu. (Russell & Norvig, 2021) Počet ćemo od dubine 1(samo korijen stabla) te postupno povećavati dubinu.

```
y_train = y_train.reshape(-1, 1)
y_test = y_test.reshape(-1, 1)

clf = DecisionTreeClassifier(max_depth=1)

clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
```

Konačno, rezultate treniranja i testiranja vizualizirat ćemo pomoću matrice konfuzije(*engl. confusion matrix).*

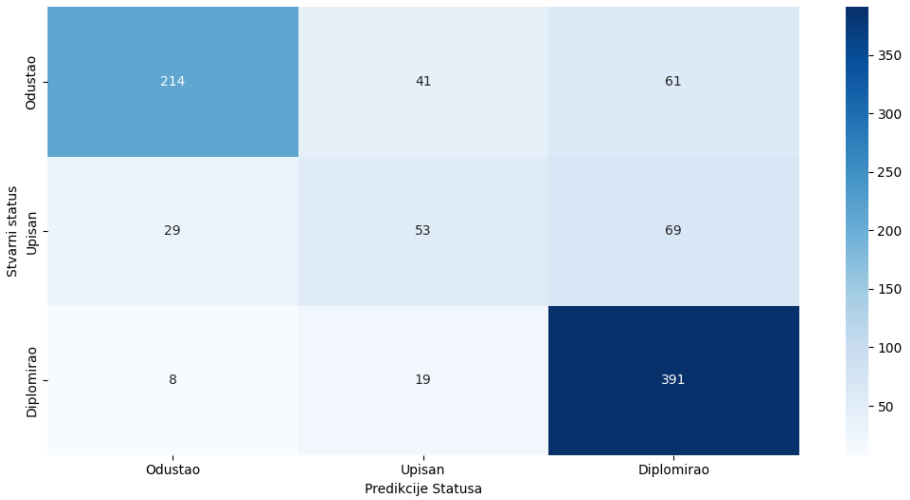
```
cm = confusion_matrix(y_test, y_pred)

plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues",
xticklabels=["Odustao", "Upisan", "Diplomirao"],
yticklabels=["Odustao", "Upisan", "Diplomirao"])
plt.xlabel("Predikcije Statusa")
plt.ylabel("Stvarni status")
plt.title("Matrica konfuzije")
plt.show()
```

Nakon treniranja modela izračunati su F1 rezultati kao metrika uspješnosti modela. Graf F1 rezultata prikazan je na slici 8.



Slika 8 Graf F1 rezultata

Najbolji F1 rezultat imalo je stablo sa vrijednosti 6 atributa `max_depth` . Predikcije tog modela prikazani su na slici 9.



Slika 9 Predikcije modela stabla odlučivanja dubinom 6

(Vlastita izrada)

4.2. Metoda k-najbližih susjeda

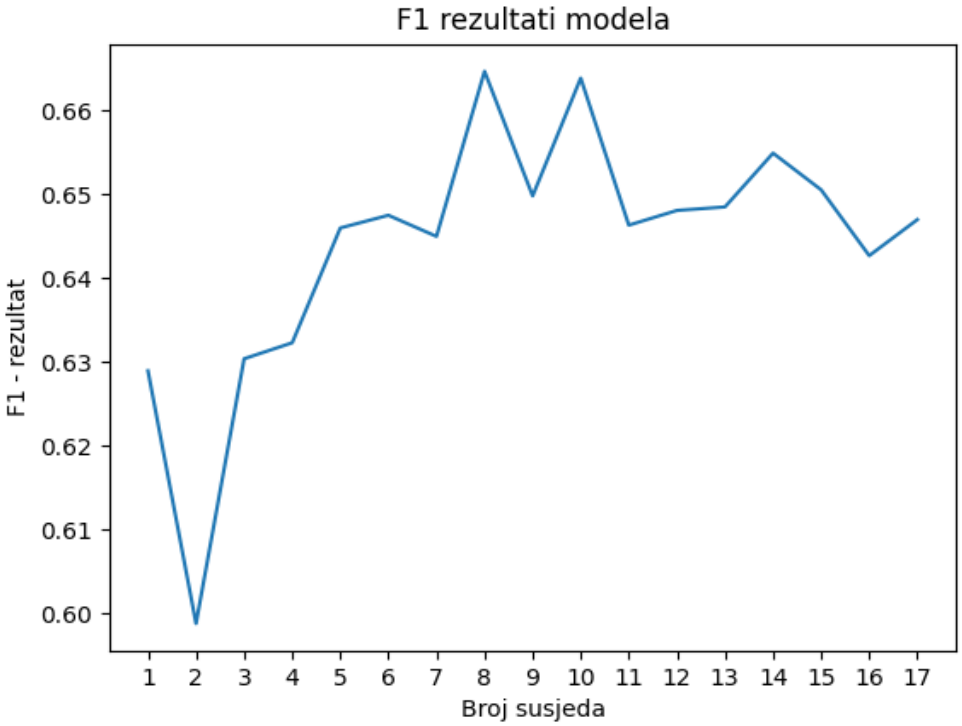
Nastavljajući se na kod koji je korišten za implementaciju stabla odlučivanja moguće je implementirati i metodu k-najbližih susjeda. Prvo je potrebno osigurati podatci koji se koriste za treniranje budu pohranjeni su u sastavljenom rasporedu memorije*(engl. contiguous memory layout)*.

```
X_train = np.ascontiguousarray(X_train)
X_test = np.ascontiguousarray(X_test)
```

Potom je još samo potrebno trenirati i testirati model. Hiperparametar pomoću kojeg ćemo optimizirati model je u ovom slučaju `n_neigbor` odnosno broj susjeda koji se uzimaju u obzir

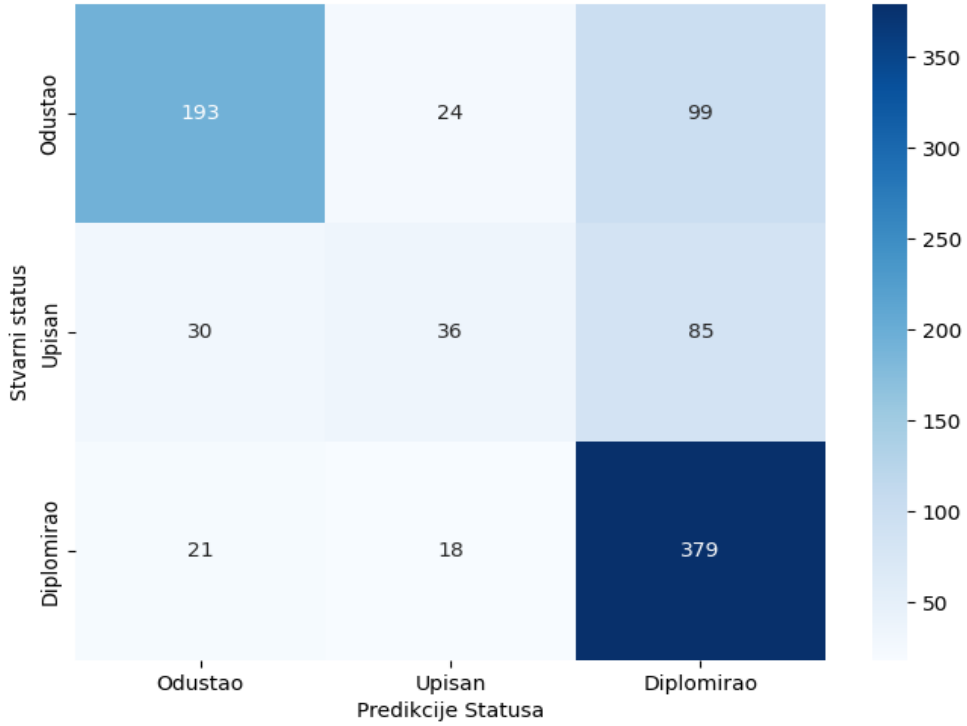
```
knn = KNeighborsClassifier(n_neighbors=i)
    knn.fit(X_train, y_train)
    y_pred = knn.predict(X_test)
```

Graf F1 rezultata prikazan je na slici 9. Na njemu je vidljivo da najbolje performanse ima model u kojem se u obzir uzima 8 najbližih susjeda. Matrica konfuzije tog modela prikazana je na slici 11.



Slika 10 Graf F1 rezultata metode K - najbližih susjeda

(Vlastita izrada)



Slika 11 Matrica konfuzije metoda K(8) - najbližih susjeda

(Vlastita izrada)

4.3. Neuronske mreže

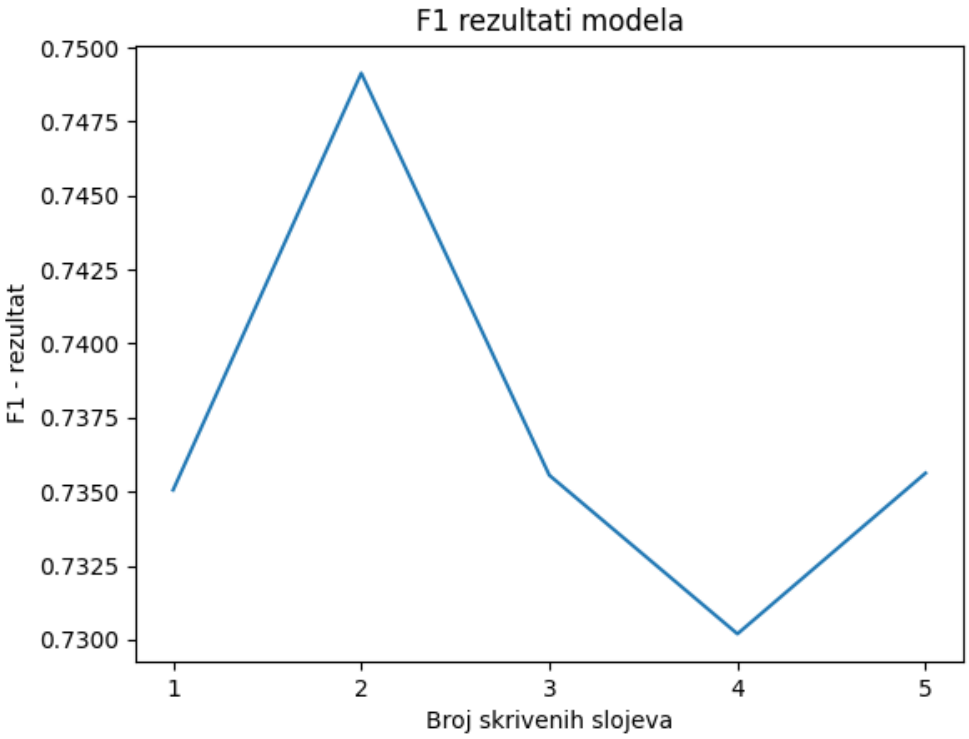
Implementacija neuronske mreže je kompleksnija od prethodnih metoda. Za implementaciju će biti korišten tensorflow 2.13.0 blioteka. Prvo je potrebno normalizirati ulazne podatke.

```
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_val = scaler.transform(X_val)
X_test = scaler.transform(X_test)
```

Potom je potrebno konstruirati arhitekturu mreže. Broj skrivenih slojeva će biti hiperparametar koji će biti optimiziran tijekom treniranja modela. Skriveni slojevi su slojevi sa oznakom `layers.Dense()`

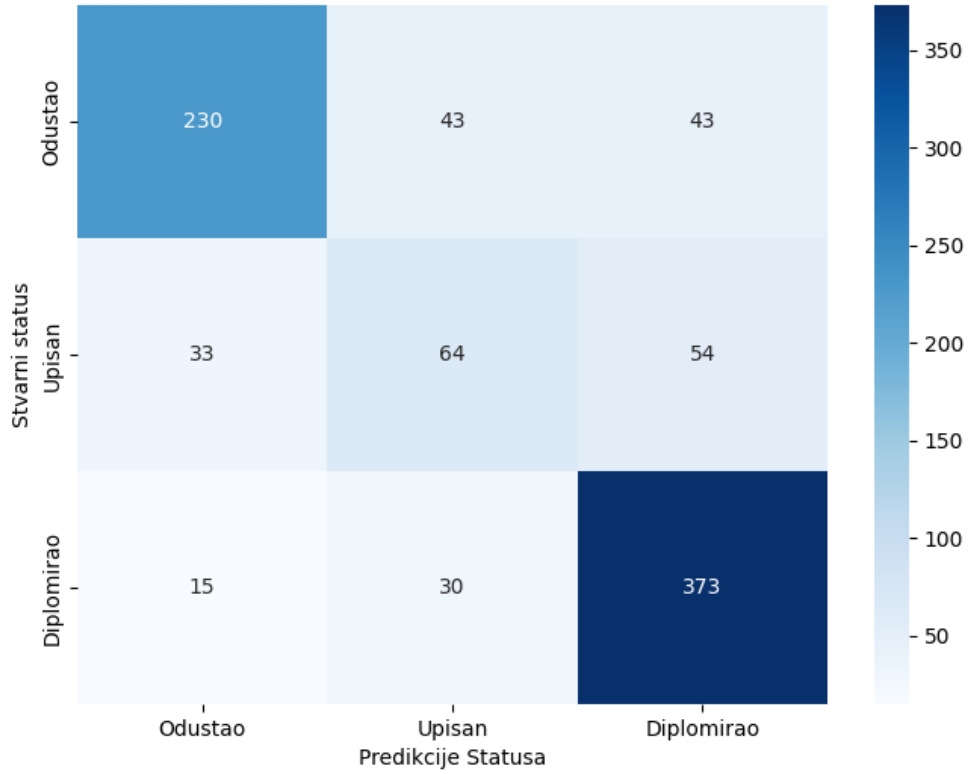
```
model = keras.Sequential([
    layers.Input(shape=(X_train.shape[1],)),
    layers.Dense(128, activation='relu'),
    layers.Dense(3, activation='softmax')
])
```

Konačno, F1 rezultati za neuronske mreže prikazani su na slici 12.



Slika 12 F1 rezultati ovisno od dubini neuronske mreže

Kao najbolji model pokazala se neuronska mreža sa dva skrivena sloja, čija je matrica konfuzije vidljiva na slici 13.



Slika 13 Matrica konfuzije neuronske mreže sa 2 skrivena sloja

4.4. Rezultati analize skupa podataka

U poglavlju 3.2 spomenuta je studija s Politehničkog instituta Portalegre koja je rezultirala skupom podataka koji će biti korišten u ovome radu. U tom je radu ukazano na puno veći broj studenata u skupu podataka koji su završili studij u skladu s propisanim trajanjem studija u donosu na broj onih koji su studiraju duže od propisanog vremena, što je vidljivo i na slici 7, stoga će vjerojatno modeli generirati predikcije bolje za studente koji su završili studij za studente koji su u skladu s propisanim trajanjem studija. Fakultet za socijalne službe i fakultet za medicinsku njegu imaju najveću stopu redovnog završetka studija, a najnižu imaju tehnološki fakulteti za proizvodnju bio-goriva te informatičko inženjerstvo. Pored ova dva, veliku stopu odustajanja od fakulteta ima i ekvikultura . U toj studiji kao najveći faktori predikcije akademske uspješnosti navedeni su podatci vezani za prva dva semestra studiranja, odabir fakulteta, vrijeme u kojima se održavaju predavanja, da li je student odselio od kuće te da li su plaćeni troškovi studiranja. (Realinho, Machado, Baptista, & Martins, 2022)

4.5. Rezultati ispitivanja uspješnosti studiranja

Nakon što su odabrani najbolji modeli svake od tehnika potrebno ih je i međusobno usporediti na validacijskom setu podataka. U kodu je potrebno samo varijable `X_test` i `y_test` zamijeniti sa varijablama `X_val` i `y_val`. Rezultati su vidljivi u tablici 3.

Tehnika strojnog učenja	F1 rezultat
Stabla odlučivanja	0.7131016670705569
K-najbližih susjeda	0.67189847785397
Neuronske mreže	0.7383731471683902

Tablica 3 Rezultati ispitivanja

Iz rezultata je vidljivo da najbolji F1 rezultat u ovom ispitivanju ima neuronska mreža, što je bilo i očekivano s obzirom da neuronske mreže uobičajeno pokazuju dobre performanse na bogatom skupu podataka.

Važno je istaknuti da iako su neuronske mreže pokazale najbolje rezultate, vrijeme implementacije, treniranja i optimizacije je također najduže kod neuronskih mreža te u scenarijima u kojima je jednostavnost od velike važnosti i druge tehnike bile vrijedne razmatranja

4.6. Interpretacija i evaluacija rezultata

S obzirom da je F1 rezultat kombinacija dvije druge metrike, ne postoji intuitivna interpretacija F1 rezultata. Također ne postoji općeprihvaćeno mišljenje o tome koji je F1 rezultat prihvatljiv, a koji ne. Jedna od mogućih interpretacija je prikazana na tablici 4.

F1 rezultat	Interpretacija
>0.9	Vrlo dobro
0.8-0.9	Dobro
0.5-0.8	OK
<0.5	Loše

Tablica 4. Interpretacija F1 rezultata

(Prema Allwright, 2022)

Vidno je da su sva tri modela u realtivno dobra te bi uz bolju optimizaciju parametara F1 rezultat bio u rasponu između 0.8 i 0.9. Ovo znači da postoji dobar balans između preciznosti i opoziva (Poglavlje 3.4) odnosno modeli su u mogućnosti precizno klasificirati većinu primjera. Na web

stranici Kaggle postoje i drugi modeli koji su trenirani na istom skupu podataka, većina modela imaju približno isti ili manji F1 rezultat od modela u ovome radu, ali postoje i neki koji imaju značajno veći F1 rezultat. Brojni modeli su koristili linearnu regresiju te su postizali F1 rezultat približno 0.85. Pored većeg F1 rezultata, modeli koji su koristili linearnu regresiju su imali kraće vrijeme treniranja te jednostavnije modele od primjerice neuronskih mreža.

Modeli trenirani u ovom radu su se pokazali kao kvalitetni te konzistentni kroz više faza testiranja. Jednostavni su za korištenje te mogu generirati stotine predikcija za manje od sekunde, ali za primjenu u stvarnom scenariju bilo bi potrebno unaprijediti arhitekturu ili koristiti drugi algoritam strojnog učenja jer nijedan od modela nije bio u mogućnosti pouzdano predvidjeti studente koji su studirali duže od predviđenog trajanja studija.

5. Zaključak

Rad se bavio usporedbom tehnika strojnog učenja u predikciji akademske uspješnosti na skupu podataka iz akademske okoline. Tehnike strojnog učenja korištene su stabla odlučivanja, metoda K-najbližih susjeda te neuronske mreže, a podatci su prikupljeni od strane znanstvenika sa Politehničkog instituta Portalegre. Za implementaciju tehnika korišten je Python 3 te implementacija sve tri metode nije zahtijevala napredno znanje programiranja. Također, sve tri tehnike su dobro dokumentirane te postoji jaka zajednica koja pruža podršku u razumijevanju tehnika.

Rezultati istraživanja su pokazali da sve tri tehnike mogu sa velikom točnošću klasificirati studente koje će završiti studij u skladu s propisanim trajanjem studija te studente koje će odustati od studija. Ovo je najvjerojatnije iz razloga što postoji značajno veći broj takvih studenata u skupu podataka. Tehnika koja se pokazala najbolja u predikciji akademske uspješnosti je tehnika neuronskih mreža, no kao što je navedeno u prethodnom poglavlju, ta tehnika je i najkompleksnija s obzirom da je potrebno dizajnirati i arhitekturu mreže.

Važno je napomenuti da prilikom optimiziranja modela nije moguće garantirati da je pronađen najbolji model. Primjerice, graf F1 rezultata neuronskih mreža je relativno nekonzistentan, što bi moglo značiti da je bila moguća bolja arhitektura mreže.

Konačno, modeli pokazuju potencijal za primjenu u stvarnom okruženju, no bilo bi potrebno još ispitivanja kako bi mogli biti sigurni. Također, poželjno bi bilo proširiti skup podataka tako da sadrži i podatke drugih sveučilišta te da je više izbalansirana zastupljenost u klasama.

Popis literature

Akçapınar, G., Altun, A., & Aşkar, P. (2019). *International Journal of Educational Techonology in Higher Education*.

Allwright, S. (22. 4 2022). *What is a good F1 score and how do I interpret it?* Preuzeto 1. 9 2023 iz Stephen Allwright: https://stephenallwright.com/good-f1-score/

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting Student Performance .

Državni zavod za statistiku. (1. 8 2022). *Državni zavod za statistiku.* Preuzeto 26. 7 2023 iz Državni zavod za statistiku: https://podaci.dzs.hr/hr/podaci/obrazovanje/

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning For Dummie.* John Wiley & Sons.

IBM. (2023). *What are neural networks?* Preuzeto 3. 8 2023 iz https://www.ibm.com/topics/neural-networks

IBM. (2023). *What is artificial intelligence?* Preuzeto 29. 7 2023 iz IBM: https://www.ibm.com/topics/artificial-intelligence

IBM. (2023). *What is Machine Learning?* Preuzeto 2. 8 2023 iz https://www.ibm.com/topics/machine-learning

Jurman, M. (2016). *Ispitivanje uspješnosti studiranja primjenom tehnika rudarenja podataka.* Varaždin: Sveučilište u Varaždinu. Preuzeto 4. 8 2023 iz https://urn.nsk.hr/urn:nbn:hr:211:850081

Kaggle. (1 2023). *Predict students' dropout and academic success*. Preuzeto 7. 8 2023 iz https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

Korstanje, J. (13. 8 2021). *The F1 score*. Preuzeto 23. 8 2023 iz Towards Data Science: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6

Ljubičić, T., & Hell, M. (2023). *Predikcija uspješnosti studenata primjenom umjetnih neuronskih mreža.*

Quinn, R. J., & Gray, G. (2019). *Prediction of student academic performance using Moodle data from a Further Education setting.* Dublin.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). *Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review.* Cáceres.

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). *Predicting Student Dropout and Academic Success.* MDPI. Dohvaćeno iz https://www.mdpi.com/2306-5729/7/11/146

Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th izd.). Pearson.

Thorn, J. (8. 3 2020). *Decision Trees Explained*. Preuzeto 23. 8 2023 iz Towards Data Science: https://towardsdatascience.com/decision-trees-explained-3ec41632ceb6

University of Helsinki & MinnaLearn. (n.d.). Elements of AI. Finska. Preuzeto 29. 7 2023 iz https://course.elementsofai.com/1/2

Wikipedia. (9. 7 2023). *Timeline of machine learning*. Preuzeto 27. 7 2023 iz Wikipedia: https://en.wikipedia.org/wiki/Timeline_of_machine_learning

Yıldırım, S. (29. 2 2020). *K-Nearest Neighbors (kNN) — Explained*. Preuzeto 23. 8 2023 iz Towards Data Science: https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3

Popis slika