DRŽAVNI UNIVERZITET U NOVOM PAZARU

TEHNIČKO-TEHNOLOŠKE NAUKE

SOFTVERSKO INŽENJERSTVO

MAŠINSKO UČENJE

GENERATIVNA VEŠTAČKA INTELIGENCIJA

Studenti:

Karišik Halida 036-024/20

Međedović Basila 036-030/20

Mentori:

prof. dr Ulfeta Marovac

doc. dr Aldina Avdić

Novi Pazar, jun 2024.

Sadržaj

Uvod

Klasifikacija podataka, kao deo mašinskog učenja, predstavlja proces automatskog kategorizovanja ili grupisanja podataka na osnovu njihovih karakteristika. U ovom radu, istražujemo primenu pet popularnih metoda mašinskog učenja u svrhu klasifikacije: logističku regresiju, Naivni Bajes, K najbližih suseda (KNN), SVM (Support Vector Machine) i stablo odlučivanja.

Logistička regresija je jedna od osnovnih tehnika za klasifikaciju podataka, posebno u binarnim klasifikacionim problemima. Naivni Bajes je probabilistička metoda koja se oslanja na pretpostavku nezavisnosti između karakteristika. KNN klasifikator koristi sličnost između podataka za klasifikaciju novih instanci. SVM je moćna tehnika koja pronalazi hiper-ravni koja najbolje razdvaja podatke različitih klasa u višedimenzionalnom prostoru. Stablo odlučivanja koristi hijerarhijsku strukturu za donošenje odluka.

Ovaj rad istražuje kako ove metode mogu biti primenjene na originalni skup podataka, kao i na sintetički generisane podatke. Prikazaćemo proces pripreme podataka, odabir relevantnih karakteristika, kao i evaluaciju performansi svake od ovih metoda korišćenjem standardnih metrika za procenu tačnosti klasifikacije.

Na kraju, cilj ovog istraživanja je ne samo da pruži uvid u primenu ovih metoda klasifikacije, već i da ilustruje njihove prednosti i ograničenja u kontekstu generativne veštačke inteligencije. Kroz analizu dobijenih rezultata, očekujemo da pružimo korisne uvide koji mogu biti od koristi u razvoju naprednih sistema veštačke inteligencije.

Klasifikacione metode

Klasifikacione metode su alati koji omogućavaju automatsko kategorizovanje ili grupisanje podataka na osnovu njihovih karakteristika. Ovi algoritmi su ključni za mnoge primene, uključujući prepoznavanje uzoraka, analizu slika, filtriranje informacija, prepoznavanje obrazaca u podacima, medicinsku dijagnostiku, prepoznavanje govora i mnoge druge oblasti.

Glavni cilj klasifikacionih metoda je naučiti model koji može predvideti klasu ili kategoriju kojoj nova instanca pripada, na osnovu njenih karakteristika ili osobina. Ovi modeli se obučavaju na osnovu skupa podataka koji sadrži instance sa već poznatim klasama, kako bi naučili vezu između ulaznih karakteristika i odgovarajućih klasa.

Logistička regresija je statistička metoda koja se često koristi za binarnu klasifikaciju podataka. Suština ove metode je modeliranje verovatnoće da određena instanca pripada određenoj klasi. Iako se naziva "regresija", logistička regresija se zapravo koristi za klasifikaciju. Model logističke regresije kombinuje ulazne karakteristike linearne funkcije sa logističkom funkcijom kako bi se izračunala verovatnoća pripadnosti instanci određenoj klasi.

Naivni Bajes je probabilistički klasifikator koji se oslanja na Bajesovu teoremu. Ključna pretpostavka ove metode je nezavisnost između karakteristika, iako to često nije realnost u stvarnim podacima. Ipak, Naivni Bajes može dati solidne rezultate, posebno u tekstualnoj analizi, kao što su klasifikacija e-pošte ili analiza sentimenta.

K najbližih suseda (KNN) je jednostavan algoritam koji se koristi za klasifikaciju i regresiju. Ova metoda radi na principu traženja K najbližih instanci iz trening skupa i pridruživanja nove instance klasi koja je najčešća među tim susedima. KNN može biti efikasan u situacijama kada su podaci dobro razdvojeni, ali može biti osetljiv na prisustvo šuma i visok broj dimenzija.

Support Vector Machine (SVM) je moćan algoritam za klasifikaciju koji pokušava pronaći hiper-ravni koja najbolje razdvaja podatke različitih klasa u višedimenzionalnom prostoru. Ova metoda koristi funkciju podrške kako bi se maksimizovala širina razmaka između različitih klasa. SVM je efikasan i daje dobre rezultate čak i u visokodimenzionalnim prostorima, ali može biti osetljiv na izbor parametara i skaliranje podataka.

Stablo odlučivanja je hijerarhijska struktura koja se koristi za donošenje odluka na osnovu serije uslova. Svaki unutarnji čvor stabla predstavlja testiranje određene karakteristike, dok listovi predstavljaju klasifikaciju. Stabla se grade iterativno, birajući karakteristiku koja najbolje razdvaja podatke na svakom koraku. Stabla odlučivanja su interpretabilna i lako ih je vizualizovati, ali mogu biti sklonija preprilagođavanju ako se ne kontroliše dubina stabla.

Opis podataka

Dataset sadrži podatke koji opisuju karakteristike radio talasa emitovanih od strane pulsara, koji stižu do Zemlje nakon prolaska kroz svemir ispunjen slobodnim elektronima. Ovi talasi obuhvataju širok spektar frekvencija, a brzina kojom se talasi usporavaju zavisi od njihove frekvencije - talasi sa višim frekvencijama se usporavaju manje u odnosu na talase sa nižim frekvencijama. Ovaj fenomen poznat je kao disperzija.

Podaci u dataset-u su organizovani u osam kolona koje sadrže sledeće informacije:

- Srednja vrednost posmatranja (Mean_Integrated): Prosečna vrednost svih posmatranih podataka.
- Ekces kurtosis posmatranja (EK): Mera koja pokazuje koliko je distribucija podataka "špicasta" ili "ravna" u odnosu na normalnu distribuciju.
- Skewness posmatranja: Mera asimetrije distribucije podataka. Pozitivna vrednost ukazuje na desnu asimetriju, dok negativna vrednost ukazuje na levu asimetriju.
- Srednja vrednost DM SNR krive (Mean_DMSNR_Curve): Prosečna vrednost krive koja prikazuje odnos disperzije merenja signala.
- Standardna devijacija DM SNR krive (SD_DMSNR_Curve): Mera varijabilnosti ili disperzije vrednosti DM SNR krive.
- Ekces kurtosis DM SNR krive (EK_DMSNR_Curve): Ekces kurtosis vrednosti DM SNR krive, koja pokazuje koliko kriva odstupa od normalne distribucije u smislu "špicatosti".
- Skewness DM SNR krive: Mera asimetrije DM SNR krive.
- Klasa (Class): Binarna oznaka koja ukazuje na to da li određeno posmatranje pripada pulsaru (1) ili ne (0).

Ovi podaci se mogu koristiti za identifikaciju pulsara, analizu njihovih karakteristika i razumevanje fenomena disperzije radio talasa u svemiru.

Priprema podataka

Balansiranje podataka postaje ključno kada imamo problem neuravnoteženosti klasa, što znači da imamo previše instanci u jednoj klasi u odnosu na drugu. Ova neuravnoteženost može dovesti do pristrasnosti modela ka dominantnoj klasi, što rezultira lošim performansama u predviđanju manjinskih klasa. Na primer, u našem slučaju imamo više podataka o elektromagnetnim zračenjima koja nisu Pulsari u odnosu na zračenja koja to jesu. Model može imati tendenciju da "pogrešno" klasifikuje elektromagnetna zračenja u većinsku klasu kako bi maksimizovao tačnost, čak i ako se zapravo radi o Pulsaru.

Balansiranje podataka omogućava modelu da ravnomerno uči iz obe klase, čime se povećava njegova sposobnost da pravilno klasifikuje manjinske klase. Ovo je posebno važno u situacijama gde je tačnost predikcije manjinskih klasa od ključnog značaja, kao što su detekcija retkih bolesti, prepoznavanje prevara u finansijskim transakcijama ili identifikacija retkih događaja u telekomunikacionim mrežama.

Bez balansiranja, modeli mogu imati tendenciju da favorizuju dominantnu klasu, što može dovesti do loše generalizacije i donošenja pogrešnih zaključaka. Stoga, balansiranje podataka omogućava bolje performanse modela i pouzdanije predikcije u realnim svetovima gde su podaci često neuravnoteženi.

Proces pripreme podataka

Za balansiranje našeg dataset-a, učitavamo podatke i odvajamo ciljnu promenljivu 'Class'. Inicijalizujemo dve nove promenljive X i Y. Varijabla X sadrži ulazne karakteristike, dok Y predstavlja ciljnu promenljivu koju želimo da predvidimo.

Da bismo primenili tehniku downsampling moramo imati uvid u tačne podatke o broju instanci obe klase.

Downsampling je tehnika za uravnoteženje skupa podataka kada postoji neuravnoteženost između klasa, što znači da jedna klasa ima značajno više instanci od druge. U kontekstu downsampling-a, ključni korak je smanjivanje broja instanci u dominantnoj klasi tako da se izjednače sa manjinskom klasom.

Nakon što izdvojimo instance za svaku klasu (class_0 i class_1) iz klase sa većim brojem instanci (u našem slučaju class_0), nasumično biramo isti broj instanci kao u klasi sa manjim brojem instanci (class_1).

Na kraju, balansirane podatke spajamo nazad u jedan skup podataka.

Evaluacija klasifikacije

Evaluacija klasifikacije je ključni korak u proceni performansi klasifikacionih modela i određivanju njihove efikasnosti u predviđanju klasa.

Tačnost meri ukupan procenat ispravno klasifikovanih instanci u odnosu na ukupan broj instanci. Ova metrika daje opštu sliku o performansama modela, ali može biti obmanjujuća u slučajevima kada imamo neuravnoteženost klasa, jer može dati visoku tačnost čak i ako model loše predviđa manjinske klase.

Odziv meri procenat tačno pozitivnih instanci među svim pozitivnim instancama u stvarnom skupu podataka. Ova metrika je korisna kada je detekcija pozitivnih instanci ključna, kao što je slučaj kod medicinskih dijagnostičkih testova ili detekcije prevara. Visok odziv znači da model dobro pronalazi pozitivne instance, ali može biti nezadovoljavajući ako dolazi do visokog broja lažno pozitivnih rezultata.

Preciznost meri procenat tačno pozitivnih instanci među svim instancama koje model klasifikuje kao pozitivne. Ova metrika se fokusira na kvalitet predviđanja pozitivnih instanci, čime se obezbeđuje da model ne daje previše lažno pozitivnih rezultata. Visoka preciznost znači da je većina pozitivnih predikcija tačna, ali može biti niska ako model propušta mnogo stvarnih pozitivnih instanci.

F-mera je harmonijski prosek između odziva i preciznosti i daje kombinovanu meru performansi modela. Ova metrika je korisna jer uzima u obzir i tačnost i potpunost klasifikacije. Visoka vrednost F-mere ukazuje na dobro balansiranje između odziva i preciznosti, što znači da model dobro predviđa obe klase.

Matrica konfuzije je tabelarna predstava performansi klasifikacionog modela koja prikazuje stvarne i predviđene klase za svaki podatak u testnom skupu. Ova matrica omogućava detaljno razumevanje performansi modela tako što razdvaja tačne i netražne predikcije za svaku klasu. Na osnovu matrice konfuzije, moguće je izračunati druge metrike poput tačnosti, odziva i preciznosti.

ROC kriva(Receiver Operating Characteristic curve) je grafikon koji ilustruje performanse binarnog klasifikatora na različitim pragovima odluke. Na x-osi je prikazan lažno pozitivni odziv (False Positive Rate), dok je na y-osi prikazan odziv (True Positive Rate). ROC kriva omogućava vizualnu procenu sposobnosti klasifikacionog modela da razlikuje između klasa i određivanje optimalnog praga odluke. Površina ispod ROC krive (AUC - Area Under the Curve) takođe pruža mjeru ukupne performanse modela, gde veća vrednost AUC-a ukazuje na bolju diskriminativnu moć modela.

GAN (Generative Adversarial Network)

U okviru savremenih istraživanja u oblasti veštačke inteligencije, Generative Adversarial Networks (GANs) su postali ključni alat za generisanje novih podataka koji su verodostojni i slični stvarnim podacima iz trening skupa. GAN predstavlja inovativnu arhitekturu neuronskih mreža koja se sastoji od dva ključna modela: **generatora i diskriminatora**, koji se takmiče u procesu učenja. Generator ima zadatak da generiše nove podatke koji su što sličniji stvarnim podacima, dok diskriminator pokušava da razlikuje između stvarnih podataka i generisanih podataka.

Tokom treninga, ova dva modela se treniraju konkurentno, što dovodi do dinamične igre gde generator postaje sve bolji u generisanju uverljivih podataka, dok diskriminator postaje sve bolji u razlikovanju stvarnih i generisanih podataka. Kada se GAN trenira dovoljno dugo, generator postaje sposoban da generiše podatke koji su toliko uverljivi da ih diskriminator ne može pouzdano razlikovati od stvarnih podataka.

U radu smo primenili Generative Adversarial Networks (GANs) koristeći dva pristupa: korišćenjem postojeće biblioteke i implementiranjem sopstvenog koda. Ovaj dvostruki pristup omogućio je dublje razumevanje koncepta GAN-a, kao i praktičnu primenu kroz već postojeća rešenja.

Prvo, iskoristili smo postojeću biblioteku za GAN, što je standardan i efikasan pristup za primenu ovog koncepta u praksi. Upotreba već razvijenih biblioteka omogućava brzu implementaciju i upotrebu visoko optimizovanog koda, što je ključno za efikasno eksperimentisanje i istraživanje različitih modela i parametara.

Drugo, kako bismo dublje razumeli unutrašnje mehanizme GAN-a, samostalno smo implementirali kod. Ovaj pristup omogućio nam je da detaljno proučimo matematičku osnovu GAN-a, kao i da prilagodimo algoritam specifičnim potrebama našeg istraživanja.

Prikaz rezultata

Tabelarni prikaz za trening skup

Metoda	Tačnost	Odziv	Preciznost	F-mera
Logistička regresija	0.788961	0.742138	0.830986	0.784053
Naivni bajes	0.750000	0.648221	0.828283	0.727273
KNN	0.817073	0.802372	0.835391	0.818548
SVM	0.786585	0.743083	0.824561	0.781705
Stablo odlučivanja	1	1	1	1

Tabelarni prikaz za test skup sa podacima iz dataset-a

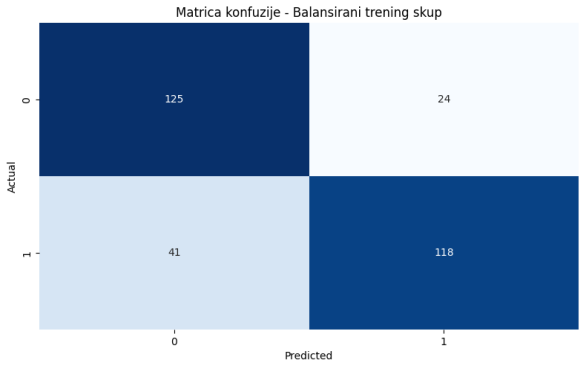
Metoda	Tačnost	Odziv	Preciznost	F-mera
Logistička regresija	0.769481	0.798658	0.770227	0.743750
Naivni bajes	0.790323	0.672727	0.822222	0.740000
KNN	0.806452	0.818182	0.762712	0.789474
SVM	0.782258	0.763636	0.750000	0.756757
Stablo odlučivanja	0.733871	0.781818	0.671875	0.722689

Tabelarni prikaz za test skup sa generisanim podacima

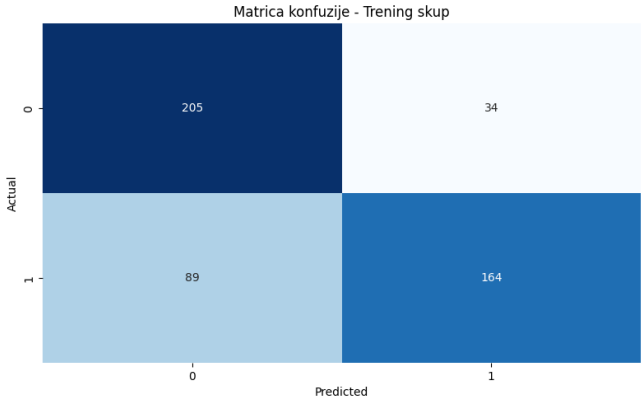
Metoda	Tačnost	Odziv	Preciznost	F-mera
Logistička regresija	0.530000	0.522472	0.382716	0.441805
Naivni bajes	0.504000	0.522472	0.363281	0.428571
KNN	0.510000	0.432584	0.348416	0.385965
SVM	0.480000	0.617978	0.364238	0.458333
Stablo odlučivanja	0.514000	0.522472	0.370518	0.433566

Matrice konfuzije za trening skup

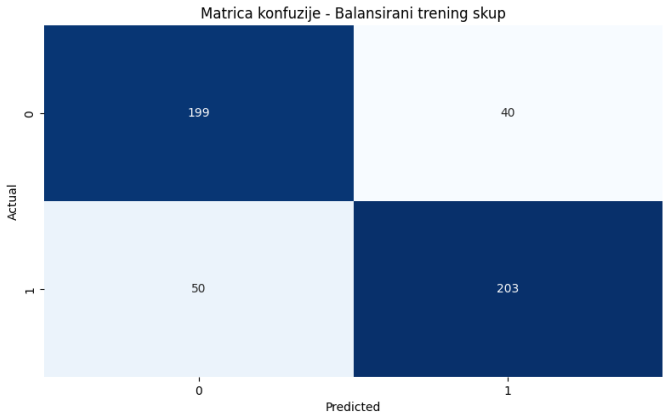
Logistička regresija



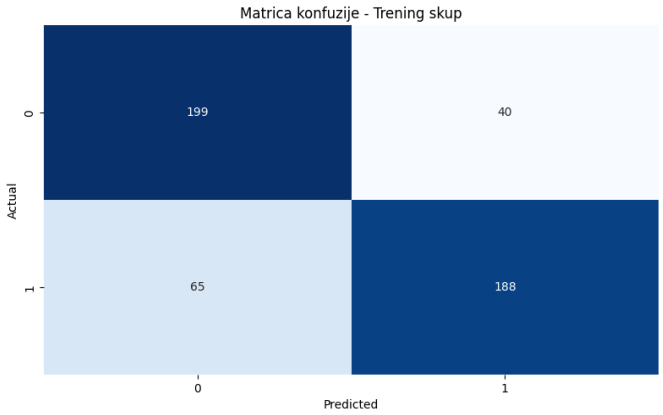
Naivni bajes



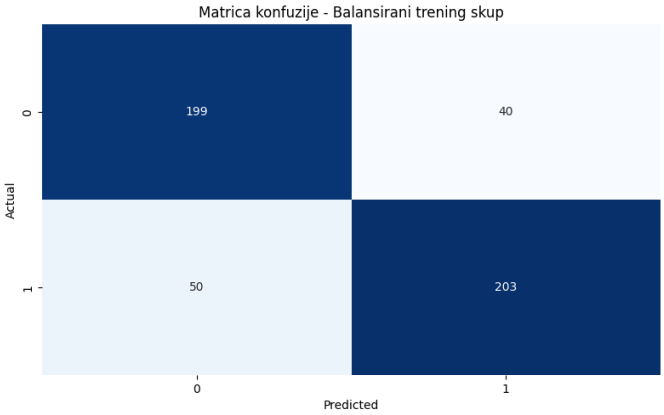
KNN



SVM



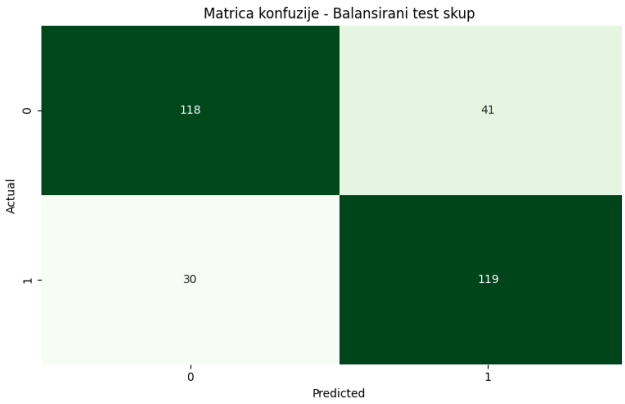
Stablo odlučivanja



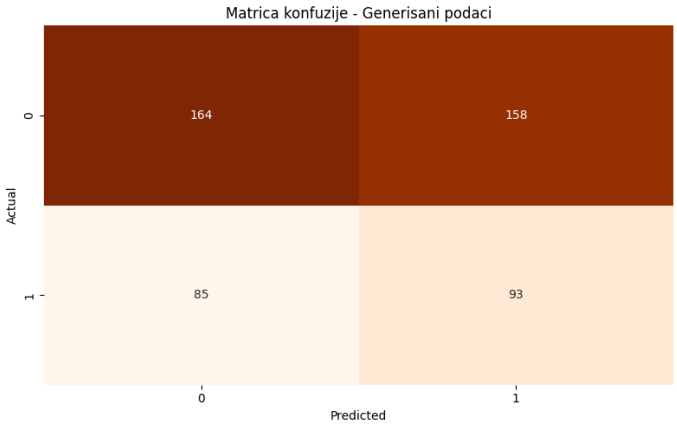
Matrice konfuzije za test skup

Logistička regresija

Regularni podaci

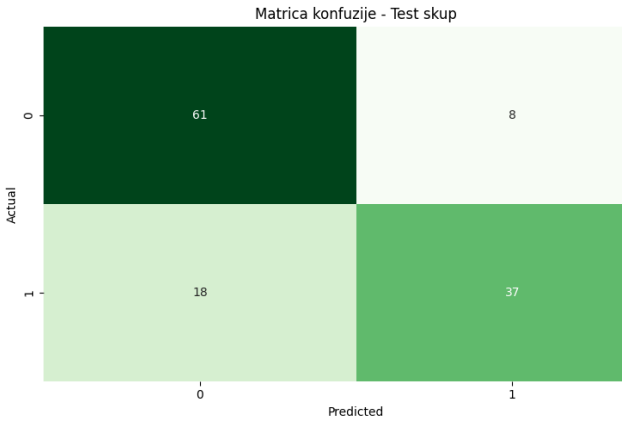


Generisani podaci

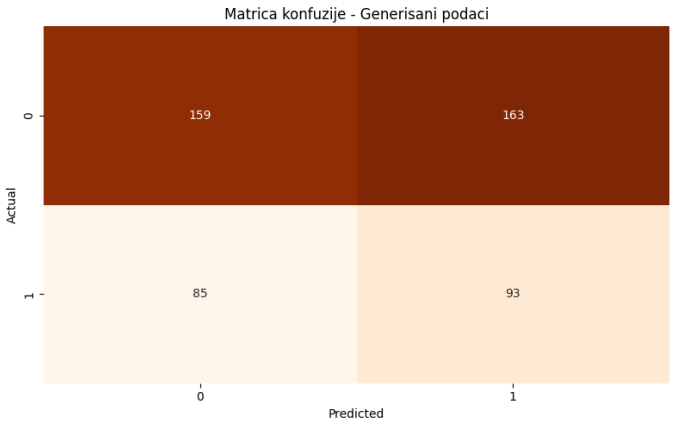


Naivni bajes

Regularni podaci

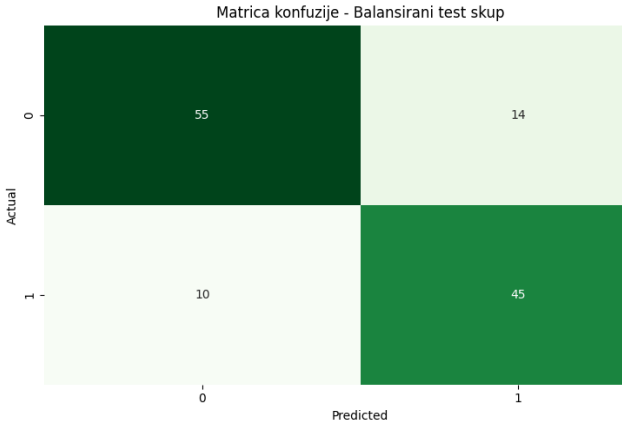


Generisani podaci

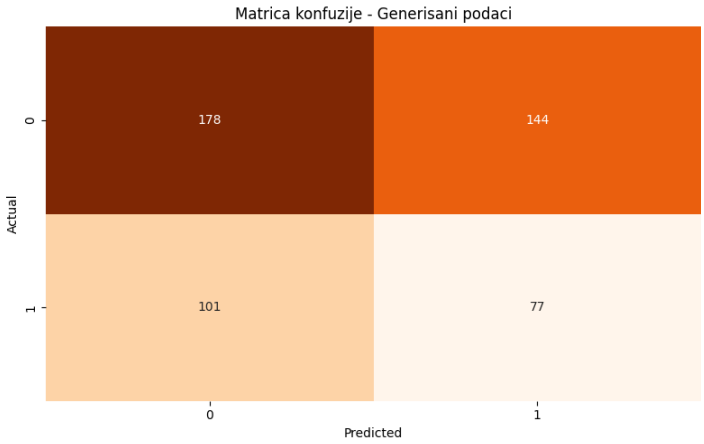


KNN

Regularni podaci

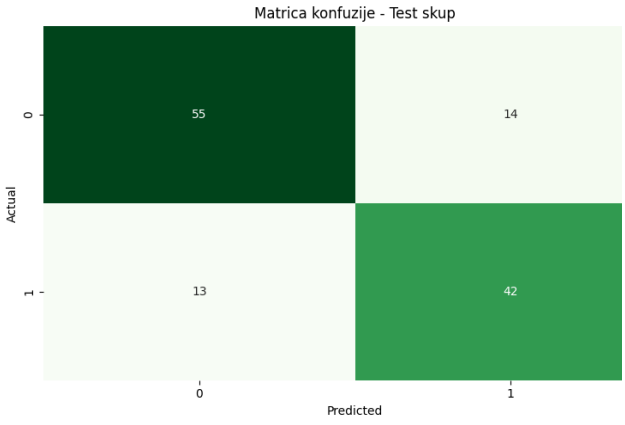


Generisani podaci

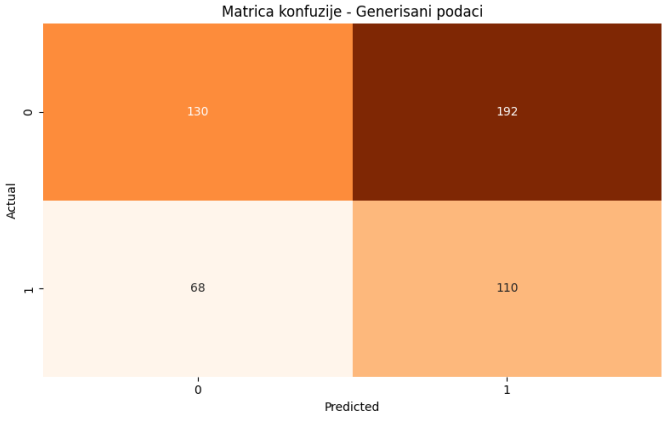


SVM

Regularni podaci

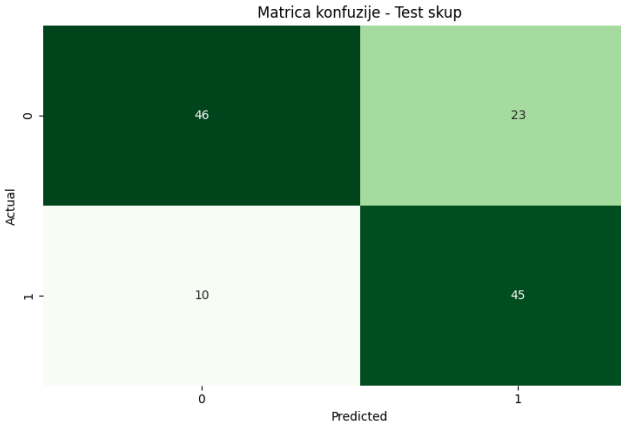


Generisani podaci

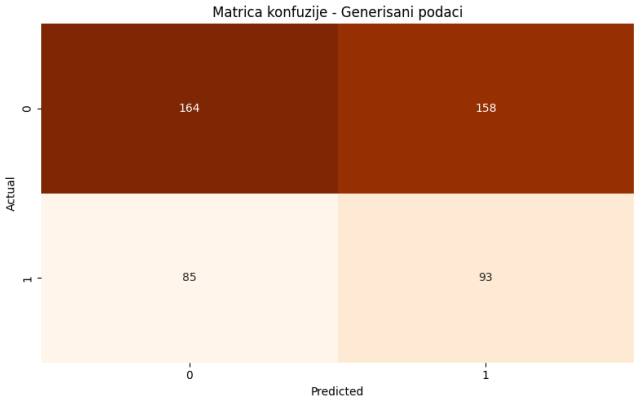


Stablo odlučivanja

Regularni podaci

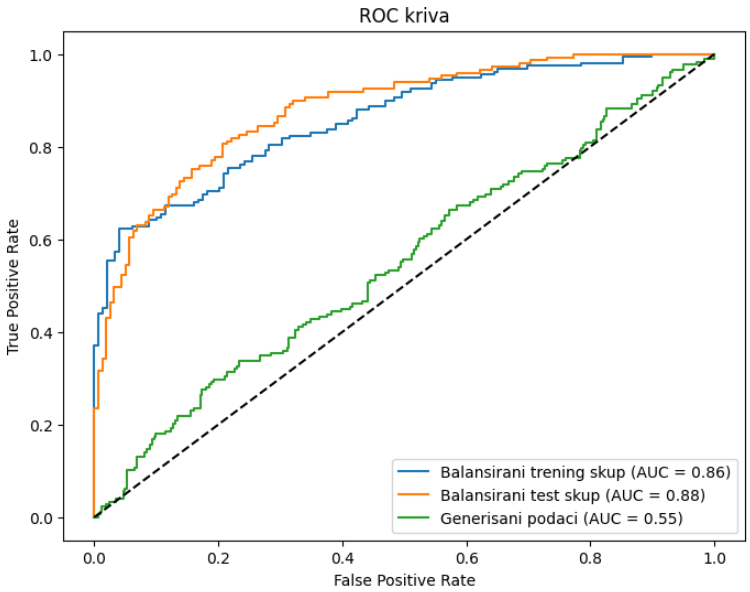


Generisani podaci

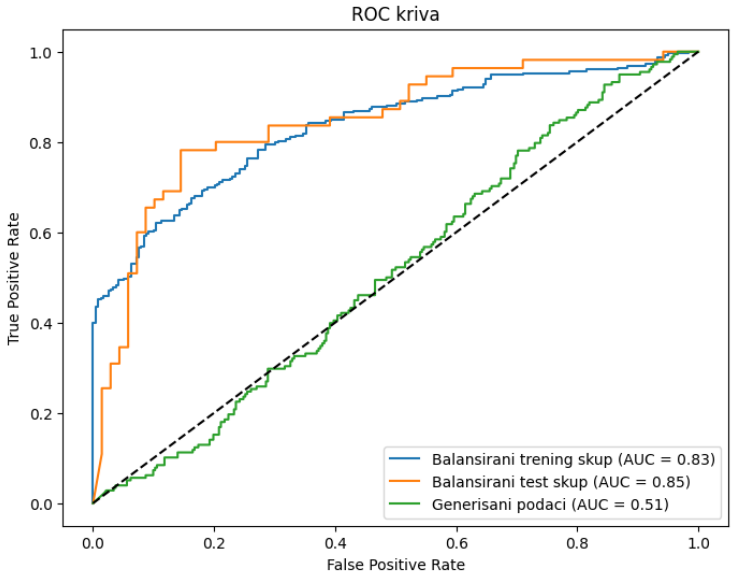


ROC krive

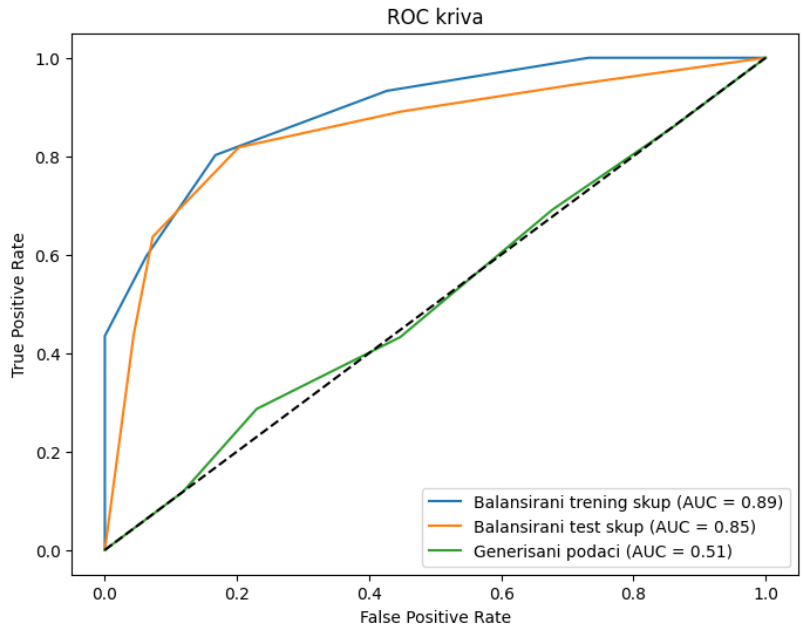
Logistička regresija



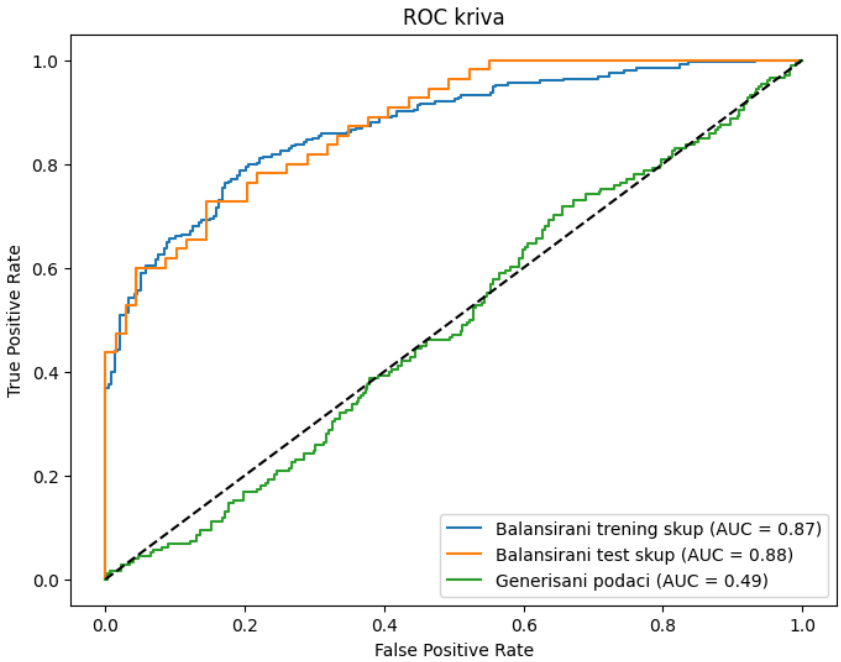
Naivni bajes



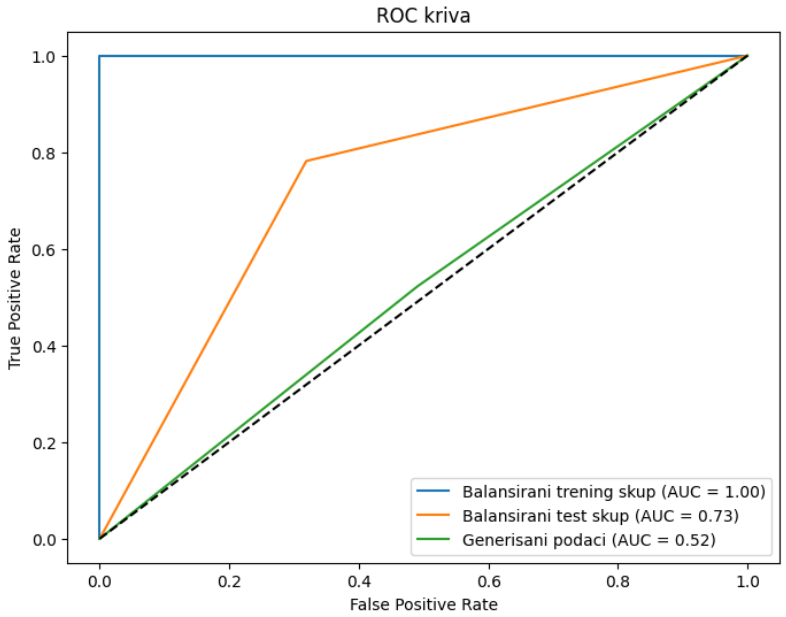
KNN


ROC kriva
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Balansirani trening skup (AUC = 0.89)
Balansirani test skup (AUC = 0.85)
Generisani podaci (AUC = 0.51)

SVM


ROC kriva
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Balansirani trening skup (AUC = 0.87)
Balansirani test skup (AUC = 0.88)
Generisani podaci (AUC = 0.49)

Stablo odlučivanja



Analiza rezultata

Na trening skupu, Stablo odlučivanja pokazuje savršene rezultate (tačnost, odziv, preciznost, i F-mera su svi 1). Ovo može ukazivati na pretreniranje, što znači da model previše dobro odgovara trening podacima, ali možda neće generalizovati dobro na nove podatke. KNN metoda pokazuje najviše performanse posle Stabla odlučivanja, sa tačnošću od 0.817073 i F-merom od 0.818548. Logistička regresija i SVM takođe pokazuju solidne performanse, dok Naivni bajes zaostaje sa nižim vrednostima.

Na test skupu sa regularnim podacima, KNN metoda ima najvišu tačnost (0.806452) i F-meru (0.789474), što ukazuje na dobru ravnotežu između odziva i preciznosti. Logistička regresija i SVM takođe pokazuju solidne rezultate, dok Naivni bajes pokazuje niži odziv. Stablo odlučivanja ima najnižu tačnost (0.733871) i F-meru (0.722689), što sugeriše da je pretreniranje na trening skupu rezultiralo slabijim performansama na test podacima.

Na test skupu sa generisanim podacima, sve metode pokazuju značajno niže performanse, što je očekivano zbog različite prirode podataka. SVM postiže najviši odziv (0.617978) i F-meru (0.458333), dok logistička regresija ima nešto bolji balans između tačnosti i F-mere. Naivni bajes i KNN pokazuju najslabije performanse u ovom slučaju.

Zaključak

Na osnovu analize rezultata, možemo izvući nekoliko važnih zaključaka o performansama različitih modela mašinskog učenja na različitim skupovima podataka. Trening skup je pokazao da Stablo odlučivanja ima savršene performanse, ali to sugeriše pretreniranje jer je malo verovatno da će takvi rezultati biti održivi na novim podacima. Nasuprot tome, KNN metoda se pokazala kao najpouzdanija sa visokim vrednostima tačnosti i F-mere, dok su logistička regresija i SVM takođe pokazali solidne rezultate.

Na test skupu sa regularnim podacima, KNN je zadržao vodeću poziciju, potvrđujući svoju sposobnost da generalizuje dobro na nove podatke. Logistička regresija i SVM su se pokazale kao konzistentne metode sa dobrim performansama. Međutim, Stablo odlučivanja je pokazalo pad u performansama, što potvrđuje problem pretreniranosti.

Test skup sa generisanim podacima otkrio je značajan pad u performansama za sve modele, što je očekivano zbog različite prirode podataka. U ovom scenariju, SVM se izdvojio sa najboljim odzivom i F-merom, što ukazuje na njegovu robusnost u rukovanju varijabilnošću u podacima. Logistička regresija je zadržala najvišu tačnost, dok su Naivni bajes i KNN pokazali najslabije performanse.

Zaključujemo da je KNN metoda najpouzdanija za regularne podatke, dok SVM pokazuje značajnu robusnost na generisanim podacima. Pretreniranje Stabla odlučivanja naglašava potrebu za daljom optimizacijom, uključujući primenu regularizacije. Dalje istraživanje i prilagođavanje modela, kao i fino podešavanje hiperparametara za KNN i SVM, preporučuje se za postizanje optimalnih rezultata u različitim uslovima.

Literatura

1. https://github.com/sdv-dev/CTGAN
2. https://www.kaggle.com/code/samanemami/gan-on-tabular-data
3. https://www.aitude.com/how-to-generate-synthetic-tabular-data-using-gan/
4. https://www.kaggle.com/datasets/prishasawhney/pulsar-classification-for-class-prediction-cleaned
5. https://www.youtube.com/watch?v=KQM6nuNsyhA
6. https://sdv.dev/SDV/user_guides/single_table/ctgan.html