Utjecaj parametara algoritma SMOTE za rukovanje problemom neuravnoteženosti klasa

Džoić, Kristian

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Sveučilišni prijediplomski studij Računarstvo

UTJECAJ PARAMETARA ALGORITMA SMOTE ZA RUKOVANJE PROBLEMOM NEURAVNOTEŽENOSTI KLASA

Završni rad

Kristian Džoić

Osijek, 2024.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA **OSIJEK**

Obrazac Z1P: Obrazac za ocjenu završnog rada na sveučilišnom prijediplomskom studiju

Ocjena završnog rada na sveučilišnom prijediplomskom studiju

Ime i prezime pristupnika:	Kristian Džoić
Studij, smjer:	Sveučilišni prijediplomski studij Računarstvo
Mat. br. pristupnika, god.	R4639, 27.07.2021.
JMBAG:	0165090946
Mentor:	doc. dr. sc. Dražen Bajer
Sumentor:	
Sumentor iz tvrtke:	
Naslov završnog rada:	Utjecaj parametara algoritma SMOTE za rukovanje problemom neuravnoteženosti klasa
Znanstvena grana završnog rada:	**Umjetna inteligencija (zn. polje računarstvo)**
Zadatak završnog rada:	Opisati problem klasifikacije s naglaskom na problem neuravnoteženosti klasa. Opisati algoritam SMOTE kao popularni algoritam za preuzorkovanje manjinske klase. Ugraditi algoritam SMOTE za potrebe eksperimentalne analize. Na nekoliko neuravnoteženih skupova podataka sustavno ispitati učinkovitost algoritma za različite postavke parametara. Rezervirano za: Kristian Džoić

Datum prijedloga ocjene završnog rada od strane mentora:	15.09.2024.
Prijedlog ocjene završnog rada od strane mentora:	Izvrstan (5)
Datum potvrde ocjene završnog rada od strane Odbora:	25.09.2024.
Ocjena završnog rada nakon obrane:	Izvrstan (5)
Datum potvrde mentora o predaji konačne verzije završnog rada čime je pristupnik završio sveučilišni prijediplomski studij:	26.09.2024.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA **OSIJEK**

IZJAVA O IZVORNOSTI RADA

Osijek, 26.09.2024.

Ime i prezime Pristupnika:	Kristian Džoić
Studij:	Sveučilišni prijediplomski studij Računarstvo
Mat. br. Pristupnika, godina upisa:	R4639, 27.07.2021.
Turnitin podudaranje [%]:	3

Ovom izjavom izjavljujem da je rad pod nazivom**: Utjecaj parametara algoritma SMOTE za rukovanje problemom neuravnoteženosti klasa**

izrađen pod vodstvom mentora doc. dr. sc. Dražen Bajer

i sumentora

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.
Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, sumentora i drugih osoba, a što je izričito navedeno u radu.

Potpis pristupnika:

SADRŽAJ

1. UVOD

Klasifikacija je jedna od ključnih metoda strojnog učenja koja služi za raspodjelu podataka u unaprijed određene kategorije ili klase. Cilj klasifikacije je razviti model koji može točno predvidjeti pripadnost novih, neviđenih podataka na temelju prethodno analiziranih uzoraka. Ova tehnika primjenjena je u mnogim područjima, uključujući medicinsku dijagnostiku, detekciju prijevara i mnoga druga polja gdje je potrebno raspodijeliti podatke po specifičnim kategorijama.

Jedan od ključnih izazova u klasifikaciji je problem neuravnoteženosti klasa (engl. *class imbalance*), koji nastaje kada su klase u skupu podataka nejednako zastupljene. Ovo nerazmjerno raspoređivanje može značajno utjecati na učinkovitost klasifikatora (engl. *classifier*), jer modeli često imaju sklonost prema predviđanju većinske klase, zanemarujući manje zastupljenu klasu. Kako bi se riješio ovaj problem, mogu se koristiti tehnike preuzorkovanja (engl. *oversampling*) i poduzorkovanja (engl. *undersampling*), koje nastoje uravnotežiti skup podataka.

Jedan od najpoznatijih i najučinkovitijih pristupa preuzorkovanju je algoritam SMOTE (engl. *synthetic minority oversampling technique*). Ovaj algoritam stvara sintetičke uzorke manjinske klase između postojećih uzoraka, čime poboljšava ravnotežu između klasa i, time, učinkovitost klasifikacije. Zbog učinkovitosti algoritma SMOTE razvijena su brojna proširenja i unaprjeđenja algoritma, ali je ipak originalni još uvijek najzastupljeniji u raznim primjenama.

Originalni SMOTE algoritam zahtijeva postavljanje dva ključna parametra prije izvođenja; broj najbližih susjeda **k** i broj sintetičkih uzoraka **q**. Ovi parametri izravno utječu na karakteristike i raspodjelu generiranih sintetičkih uzoraka i, posljedično, na učinkovitost klasifikacije. Budući da ne postoje univerzalno dobre postavke ovih parametara, potrebno ih je prilagoditi specifičnostima svakog pojedinog problema.

U drugom poglavlju ovog rada detaljno je razrađen pojam klasifikacije i problem neuravnoteženosti klasa. Predstavljeni su različiti klasifikacijski modeli te načini evaluacije njihovih rezultata. Također, detaljno je objašnjen algoritam SMOTE, uključujući njegove prednosti i nedostatke, kao i njegove uobičajene postavke parametara. Treće poglavlje fokusira se na detaljan opis programskog rješenja razvijenog za implementaciju SMOTE algoritma. Opisani su svi ključni aspekti i komponente implementacije, uključujući funkcionalnosti i način rada aplikacije. U četvrtom poglavlju provedena je eksperimentalna analiza primjene SMOTE algoritma s različitim postavkama parametara. Ovdje su predstavljeni rezultati testiranja i

usporedba učinkovitosti različitih konfiguracija SMOTE-a u rješavanju problema neuravnoteženosti klasa.

1.1. Zadatak završnog rada

Opisati problem klasifikacije s naglaskom na problem neuravnoteženosti klasa. Opisati algoritam SMOTE kao popularni algoritam za preuzorkovanje manjinske klase. Ugraditi algoritam SMOTE za potrebe eksperimentalne analize. Na nekoliko neuravnoteženih skupova podataka sustavno ispitati učinkovitost algoritma za različite postavke parametara.

2. PROBLEM NEURAVNOTEŽENOSTI KLASA I ALGORITAM SMOTE

Klasifikacija je tehnika nadziranog strojnog učenja u kojoj model nastoji pronaći povezanost između vrijednosti značajki primjera $\boldsymbol{x}$ i pripadajuće klase $\boldsymbol{y}$ kojom je taj primjer označen, s ciljem predviđanja oznaka za nove, nepoznate primjere [1]. Ako oznaka $\boldsymbol{y}$ može poprimiti samo dvije vrijednosti, riječ je o binarnoj klasifikaciji, dok se kod više mogućih oznaka radi o višeklasnoj klasifikaciji.

Kada je broj uzoraka jedne klase značajno veći od broja uzoraka druge klase nastaje pojava poznata kao neuravnoteženost klasa. Ovakva situacija najčešće dovodi do pristranosti klasifikatora prema većinskoj klasi (engl. *majority class*), uz istovremeno zanemarivanje manjinske klase (engl. *minority class*). To može značajno umanjiti učinkovitost i korisnost modela klasifikacije.

2.1. Klasifikacija i problem neuravnoteženosti klasa

Klasifikacija podrazumijeva raspoređivanje primjera u pripadajuće klase na temelju vrijednosti značajki koje ih opisuju [2, 3]. Kao što je ranije navedeno, ovisno o broju klasa postoje binarna i višeklasna klasifikacija. U ovom radu fokus će biti na primjerima binarne klasifikacije s podacima iz skupova $\mathcal{A} \subset \mathbb{R}^m$, gdje su značajke koje opisuju podatke u obliku realnih vrijednosti.

Klasifikacija ima brojne različite primjene u raznim djelatnostima. Na primjer, u strojnom vidu, sustav analizira slike koje dobiva putem kamere. Tipična primjena takvog sustava je u industrijskoj proizvodnji za vizualnu kontrolu ili automatizaciju proizvodnih linija. U procesu kontrole, sustav određuje je li proizvedeni predmet u skladu sa specifikacijama, odnosno, utvrditi postoje li neke nepravilnosti ili ne [3]. Jedna od češćih primjena modela klasifikacije je u medicinskoj dijagnostici. Klasifikacijski modeli mogu se koristiti za procjenu prisutnosti neke bolesti kod pacijenta na temelju različitih nalaza [4]. Također, ovi modeli mogu pomoći i pri raspoređivanju tumora na zloćudne i dobroćudne, što može značajno doprinijeti ranoj detekciji različitih vrsta raka [5]. Osim navedenih primjera, klasifikacijski modeli nalaze široku primjenu u mnogim drugim područjima. Na primjer, koriste se za automatsko prepoznavanje glasa, prepoznavanje otisaka prstiju, prepoznavanje DNA sekvenci i slične zadatke [2].

U brojnim primjenama nadziranog strojnog učenja postoji značajna razlika u vjerojatnostima pojavljivanja različitih klasa, odnosno vjerojatnost da primjer pripada jednoj klasi može biti znatno veća nego da pripada drugoj. Ova pojava, poznata kao neuravnoteženost klasa, često se

javlja u raznim područjima poput telekomunikacija, financija, biologije, medicine i brojnih drugih. Neuravnoteženost klasa najčešće rezultira pristranošću klasifikatora prema većinskoj klasi. Glavni razlozi za to su činjenica da su pravila klasifikacije koja predviđaju manjinsku klasu često visoko specijalizirana i imaju vrlo nisku pokrivenost, zbog čega se odbacuju u korist općenitijih pravila, odnosno onih koja predviđaju većinsku klasu. Također male skupine podataka manjinske klase često se svrstaju u šum te ih stoga često klasifikator zanemaruje. Nadalje, vrijedi naglasniti da je upravo manjinska klasa često od većeg interesa u mnogim primjenama te stoga njena netočna klasifikacija može predstavljati značajan problem [6].

Problemom neuravnoteženosti klasa može se rukovati na više različitih načina. Oni se mogu podijeliti u dvije kategorije: pristupi temeljeni na algoritmima i pristupi temeljeni na podacima. Pristupi temeljeni na algoritmima uključuju dodavanje dodatnih pravila u klasifikatore, poput posebnih kaznenih funkcija po klasi ili uvođenja pristranosti u učenju. Pristupi temeljeni na podacima oslanjaju se isključivo na podatke te koriste poduzorkovanje, preuzorkovanje ili njihove kombinacije, bilo nasumično ili temeljeno na svojstvima uzoraka. Prednost pristupa temeljenih na podacima je njihova fleksibilnost i jednostavnost, zbog čega se često koriste [7].

Poduzorkovanje podrazumijeva uklanjanje primjera većinske klase iz skupa podataka kako bi se smanjila razlika između broja primjera većinske i manjinske klase. Preuzorkovanje, s druge strane, odnosi se na povećanje broja primjera manjinske klase kako bi se postigla bolja uravnoteženost. Ovo se može postići nasumičnim dupliciranjem primjera manjinske klase ili generiranjem novih primjera temeljenih na postojećim. Preuzorkovanje se često preferira u odnosu na poduzorkovanje zobg izražene mogućnosti gubitka bitnih uzoraka većinske klase kod poduzorkovanja. Ipak, u nekim slučajevima i na određenim skupovima podataka, poduzorkovanje može dovesti do boljih performansi klasifikatora, te stoga ne postoji jasan konsenzus koji pristup daje bolje rezultate [7].

2.1.1. Neki popularni algoritmi za klasifikaciju

Postoji mnogo različitih algoritama za klasifikaciju, odnosno klasifikatora, koji se mogu primijeniti na različite probleme u strojnom učenju. Međutim, nisu svi klasifikatori jednako prikladni za sve probleme. Odabir pravog klasifikatora (engl. *model selection*) predstavlja zaseban i važan problem u strojnom učenju [2]. Prikladan klasifikator može značajno poboljšati performanse, dok neodgovarajući klasifikator može dovesti do loših rezultata i male učinkovitosti.

2.1.1.1. Algoritam *k*-najbližih susjeda

Algoritam *k*-najbližih susjeda (engl. *k-nearest neighbors*, *k*-NN) je klasifikacijski model čiji se princip rada zasniva na određivanju klase primjera na temelju klasa njegovih *k* najbližih susjeda. Udaljenost između podataka, koja se koristi za identifikaciju najbližih susjeda, određuje se prema određenoj mjeri udaljenosti, pri čemu se najčešće koristi Euklidska udaljenost [2]. Osim postavljanja mjere udaljenosti, algoritmu *k*-NN može se postaviti i parametar *k*, koji označava broj susjeda koji se uzimaju u obzir pri klasifikaciji primjera.

Neke od najvažnijih prednosti *k*-NN algoritma su jednostavnost implementacije te činjenica da ne zahtijeva fazu treniranja zato što koristi cijeli skup podataka za klasifikaciju. Međutim, postoje i značajni nedostatci ovog algoritma, među kojima je najizraženija potreba za računanjem udaljenosti između testnog primjera i svih primjera u skupu podataka te pohrana tih udaljenosti u računalnoj memoriji, što može biti vrlo zahtjevno na velikim skupovima podataka.

2.1.1.2. Gaussov naivni Bayes klasifikator

Gaussov naivni Bayes klasifikator (engl. *Gaussian Naive Bayes classifier*, GNB) je klasifikacijski model zasnovan na Bayesovom teoremu i pretpostavci o nezavisnosti značajku unutar svake klase. Ovaj algoritam pretpostavlja da su značajke distribuirane prema Gaussovoj (normalnoj) distribuciji. Za klasifikaciju koristi priorne vjerojatnosti za svaku klasu, kao i srednje vrijednosti i standardne devijacije značajki unutar svake klase. Ove vrijednosti predstavljaju parametre modela i računaju se iz podskupa za treniranje [3].

Prednosti GNB klassifikatora su jednostavnost implementacije te brzina treniranja modela. Dobro funkcionira s velikim brojem značajki i velikim skupovima podataka. Međutim, značajan nedostatak ovog modela je pretpostavka o nezavisnosti značajki, koja često nije stvarna i može negativno utjecati na točnost klasifikacije. Također, moguće je i smanjenje preciznosti modela ukoliko značajke nisu normalno distibuirane.

2.1.1.3. Stablo odluke

Stablo odluke (engl. *Decision Tree*, DT) je klasifikacijski model koji koristi strukturu stabla za provođenje klasifikacije. Ovaj model radi na principu donošenja odluka temeljenih na značajkama podataka, gdje svaki čvor u stablu predstavlja provjeru vrijednosti određene značajke, a različite grane koje izlaze iz čvora odgovaraju mogućim vrijednostima te značajke. Listovi stabla predstavljaju konačne klase koje je model predvidio [2]. Za izgradnju stabla koristi

se kriterij za odabir najboljih značajki za podjelu podataka, poput *Gini* indeksa ili entropije. Stablo se gradi rekurzivno tako da se podaci dijele na temelju odabrane značajke sve dok se ne postignu određeni kriteriji za zaustavljanje, poput maksimalne dubine stabla.

Prednosti stabla odluke uključuju jednostavnost vizualizacije rezultata te sposobnost obrade podataka različitih vrsta i struktura. Međutim, stabla odluke su često sklona prekomjernom prilagođavanju (engl. *overfitting*) podacima nad kojim se treniraju, posebno ako je stablo previše duboko [3].

2.1.1.4. Stroj s potpornim vektorima

Stroj s potpornim vektorima (engl. *Support Vector Machine,* SVM) je klasifikacijski model koji se temelji na pronalasku optimalne hiperravnine koja razdvaja podatke u različite klase. Princip rada algoritma zasniva se na maksimizaciji margine, odnosno udaljenosti između hiperravnine i najbližih uzoraka svake klase, koji se nazivaju potporni vektori (engl. *support vectors)* [2]. SVM se često koristi za rješavanje problema klasifikacije u prostoru visoke dimenzionalnosti, a može se koristiti i za nelinearne probleme. Nelinearni podaci mogu se transformirati u višu dimenziju pomoću funkcija jezgre (engl. *kernel*) gdje je moguće pronaći linearnu hiperravninu koja razdvaja klase. Najčešće korištene funkcije jezgre uključuju linearne, polinomske, sigmoid te radijalne funkcije jezgre (engl. *radial basis function,* RBF). Parametri SVM-a uključuju parametar *C,* koji kontrolira ravnotežu između maksimiziranja margine i minimiziranja grešaka klasifikacije, te parametre funkcije jezgre.

Prednosti SVM-a uključuju njegovu sposobnost da se nosi s visokodimenzionalnim podacima i pronađe optimalnu hiperravninu, što često dovodi do visoke točnosti klasifikacije. Međutim, nedostatci SVM-a uključuju visoke računalne zahtjeve u slučaju velikih skupova podataka i potrebu za pažljivim odabirom parametara koji dosta utječu na performanse modela. Također, na velikim skupovima podataka s velikim brojem uzoraka i značajki, učinkovitost SVM-a može biti znatno smanjena.

2.1.2. Vrednovanje učinkovitosti algoritama klasifikacije

Nakon što je izvršena klasifikacija primjera iz podskupa za testiranje, uspoređuju se predviđene klase od strane klasifikatora s njihovim stvarnim klasnim oznakama iz skupa podataka. Za prikaz rezultata te usporedbe najčešće se koristi matrica zbunjenosti (engl. *confusion matrix*), posebice u slučajevima binarne klasifikacije. Matrica zbunjenosti za binarnu klasifikaciju ima oblik tablice s dva retka i dva stupca [1].

Elementi takve matrice su:

- Istinito negativni (engl. *True negative*, TN) – primjeri koji su točno klasificirani kao negativni
- Istinito pozitivni (engl. *True positive*, TP) – primjeri koji su točno klasificirani kao pozitivni
- Lažno negativni (engl. *Flase negative*, FN) – primjeri koji su netočno klasificirani kao negativni
- Lažno pozitivni (engl. *False positive*, FP) – primjeri koji su netočno klasificirani kao pozitivni

Elementi na glavnoj dijagonali matrice zbunjenosti prikazuju broj točno klasificiranih primjera iz testnog skupa, dok elementi izvan glavne dijagonale prikazuju netočno klasificirane primjere. Standardni oblik matrice zbunjenosti za binarnu klasifikaciju prikazan je tablicom 2.1.

Tablica 2.1. Matrica zbunjenosti binarne klasifikacije

	Predviđeno negativni	**Predviđeno pozitivni**
Stvarno negativni	Istinito negativni (TN)	Lažno pozitivni (FP)
Stvarno pozitivni	Lažno negativni (FN)	Istinito pozitivni (TP)

Jedna od najčešćih mjera za procjenu učinkovitosti klasifikatora je točnost klasifikacije (engl. *classification accuracy*). Ova mjera predstavlja omjer točno klasificiranih primjera u odnosu na ukupan broj primjera koji su klasificirani te se računa prema formuli:

$$Točnost = \frac{TN + TP}{TN + TP + FN + FP} \tag{2-1}$$

Točnost klasifikacije je jednostavna mjera koja može dati dobar uvid u učinkovitost klasifikatora. Međutim, kod neuravnoteženih podataka ova mjera više nije prikladna zato što ne pravi razliku između broja točno klasificiranih primjera različitih klasa [6]. Na primjer, u skupu s velikim omjerom neuravnoteženosti, točnost klasifikacije može doseći vrijednosti bliske 1 ako klasifikator sve primjere svrsava u većinsku klasu. Zbog toga točnost klasifikacije može zavarati i pružiti lažan dojam uspješnosti modela, iako zapravo ignorira manjinsku klasu, čija je klasifikacija od najvećeg interesa.

Tako se zbog ograničenja mjere točnosti klasifikacije za precizniju procjenu učinkovitosti klasifikatora na neuravnoteženim podacima, koriste mjere koje bolje odražavaju uspješnost modela u prepoznavanju obje klase. Dvije važne mjere u ovom kontekstu su F-1 mjera (engl. *F-1 score*) i geometrijska sredina istina (engl. *geometric mean of trues,* G-Mean).

F-1 mjera kombinira preciznost (engl. *precision*) i odziv (engl. *recall*) u jednu mjeru koja pruža uravnoteženu procjenu učinkovitosti modela [1]. Preciznost je definirana kao omjer točno klasificiranih pozitivnih primjera u odnosu na sve primjere koji su klasificirani kao točni te se računa prema formuli:

$$Preciznost = \frac{TP}{TP + FP} \tag{2-2}$$

Odziv predstavlja omjer točno klasificiranih pozitivnih primjera u odnosu na sve stvarno pozitivne primjere te se računa prema formuli:

$$Odziv = \frac{TP}{TP + FN} \tag{2-3}$$

F-1 mjera predstavlja harmonijsku sredinu preciznosti i odziva te je definirana formulom:

$$F_1 = 2 \times \frac{Preciznost \times Odziv}{Preciznost + Odziv} \tag{2-4}$$

Ova mjera je korisna zato što pruža sveobuhvatan uvid u performanse modela u prepoznavanju pozitivne klase, posebice kada je važna ravnoteža između točnosti i potpunosti prepoznavanja pozitivnih primjera.

G-Mean mjera uzima u obzir odziv za pozitivnu klasu te specifičnost za negativnu klasu, a računa se kao geometrijska sredina ove dvije mjere [1]. Specifičnost (engl. *specificity*) je definirana kao omjer točno klasificiranih negativnih primjera u odnosu na ukupan broj stvarno negativnih primjera:

$$Specifičnost = \frac{TN}{TN + FP} \tag{2-5}$$

G-Mean kombinira odziv i specifičnost na sljedeći način:

$$G\text{-}Mean = \sqrt{Odziv \times Specifičnost} \tag{2-6}$$

Ova mjera pomaže u održavanju ravnoteže između uspješnosti modela u prepoznavanju obje klase, čime se smanjuje utjecaj neuravnoteženosti klasa i pruža potpunija slika o ukupnoj učinkovitosti klasifikatora.

Korištenjem ovih mjera umjesto točnosti klasifikacije, moguće je dobiti bolji uvid u uspješnost klasifikacijskog modela u situacijma s neuravnoteženim podacima, gdje su ravnoteža i učinkovitost prepoznavanja obje klase ključni za pravilnu procjenu korisnosti modela.

2.2. **Algoritam SMOTE**

Kao što je prethodno spomenuto neuravnoteženost skupa podataka u pravilu rezultira smanjenom učinkovitošću klasifikatora. Ta smanjena učinkovitost je posljedica pristranosti klasifikatora većinskoj klasi. Često se kao način smanjenja neuravnoteženosti skupa i, samim time poboljšanja učinkovitosti klasifikatora koriste algoritmi preuzorkovanja. Kod preuzorkovanja neuravnoteženog skupa podataka postoje dva glavna pristupa:

- Nasumično preuzorkovanje – ova metoda preuzorkovanja temelji se na dupliciranju uzoraka manjinske klase dok se ne dostigne željeni omjer neuravnoteženosti. Uzorci manjinske klase koji se dupliciraju biraju se nasumično [8]. Primjena nasumičnog preuzorkovanja često rezultira prekomjernim prilagođavanjem klasifikatora dupliciranim uzorcima, bez značajnog poboljšanja u performansama modela [7]. Pseudo-kod ovog algoritma prikazan je slikom 2.1.
- SMOTE – ovaj pristup razvijen je kao alternativa nasumičnom preuzorkovanju te se temelji na generiranju sintetičkih uzoraka manjinske klase koji su slični izvornim uzorcima, ali ne identični. Manjinska klasa se preuzorkuje tako što se za već postojeći uzorak manjinske klase generiraju sintetički uzorci duž linijskih segmenata koji spajaju taj uzorak s njegovim najbližim susjedima iz manjinske klase [9].

Parametri algoritma SMOTE su :

- **k** – cjelobrojni parametar koji označava broj najbližih susjeda koji se koriste za generiranje sintetičkih uzoraka. Odabir ovog parametra utječe na razinu varijabilnosti generiranih uzoraka.
- **q** – cjelobrojni parametar koji označava broj novih sintetičkih uzoraka koje treba generirati za svaki postojeći uzorak manjinske klase. Ovaj parametar određuje koliko će se izvorni skup podataka proširiti novim uzorcima manjinske klase.

SMOTE algoritam djeluje na način da se za svaki uzorak manjinske klase **x** odredi skup njegovih **k** najbližih susjeda iz manjinske klase, pri čemu se mogu koristiti različite metričke funkcije (najčešće se koristi Euklidska udaljenost). Zatim se za svaki uzorak **x** generira se **q** sintetičkih uzoraka prema formuli:

$$s^i := x + U_i(0,1) \times (x^{r(i)} - x) \tag{2-7}$$

gdje je $U_i(0,1)$ uniformna varijabla u rasponu (0,1), a $x^{r(i)}$ nasumično odabrani susjed uzorka **x** [7]. Pseudo-kod SMOTE algoritma prikazan je na slici 2.2.

```
Nasumično preuzorkovanje(n)
1:  ZA i=1 DO n RADI
2:      Nasumično odaberi uzorak manjinske klase
3:      Spremi značajke odabranog uzorka
4:      Dodaj novi uzorak sa spremljenim značajkama
5:  KRAJ
```

Slika 2.1. Pseudo-kod nasumičnog preuzorkovanja

```
SMOTE(k,q)
1: ZA i = 1 DO broj_uzoraka RADI
2:      Odredi k najblizih susjeda za Xi i spremi ih u N_niz
3:      ZA j = 1 DO q RADI
4:          Nasumicno odaberi susjeda n iz N_niz
5:          sint = novi uzorak
6:          ZA znacajka = 1 DO broj_znacajki RADI
7:              d = udaljenost izmedu Xi_znacajka i n_znacajka
8:              U = nasumican broj izmedu 0 i 1
9:              sint_znacajka = Xi_znacajka + U * d
10:         KRAJ
11:         Dodaj sint u skup
12:     KRAJ
13: KRAJ
```

Slika 2.2. Pseudo-kod algoritma SMOTE

2.2.1. Prednosti i nedostatci

Glavna prednost algoritma SMOTE je u tome što uspješno rješava problem neuravnoteženosti klasa generiranjem sintetičkih uzoraka manjinske klase, čime smanjuje pristranost klasifikatora prema većinskoj klasi i u većini slučajeva poboljšava njegovu učinkovitost na neuravnoteženim skupovima podataka. SMOTE ima prednost nad nasumičnim poduzorkovanjem jer smanjuje rizik gubitka korisnih uzoraka većinske klase. U odnosu na nasumično preuzorkovanje, SMOTE se u većini slučajeva pokazuje bolji jer je manje sklon prekomjernom prilagođavanju klasifikatora podskupu za treniranje [10].

Jedan od izraženijih nedostataka SMOTE algoritma je prekomjerna generalizacija (engl. *overgeneralization*), jer algoritam generira sintetičke uzorke bez obzira na distribuciju većinske klase, što može dovesti do neprecizne klasifikacije. Također, SMOTE može povećati rizik preklapanja klasa u situacijama kada su primjeri manjinske klase rijetko raspoređeni, jer se sintetički uzorci mogu nalaziti u područjima koja pripadaju većinskoj klasi, što dodatno otežava razlikovanje između klasa [11].

Također, značajan nedostatak SMOTE algoritma je i uvođenje sintetičkih primjera, što može biti teško opravdati, posebno u područjima poput medicine. U takvim domenama, stvaranje sintetičkih primjera može izazvati zabrinutost u vezi valjanosti rezultata, jer umjetno generirani podaci, poput "sintetičkih pacijenata", nemaju smisla za donošenje praktičnih odluka [10].

SMOTE algoritam često koristi jednostavne mjere za izračunavanje udaljenosti između uzoraka, što može predstavljati problem kada se radi s podacima koji imaju svojstvo mnogostrukosti (engl. *manifold*). Kod takvih podataka, stvarna distribucija podataka nije ravnomjerno raspoređena u cijelom prostoru značajki, već je koncentrirana u niže-dimenzionalnom prostoru unutar originalnog prostora značajki. Primjena SMOTE-a na ovim podacima može rezultirati sintetičkim uzorcima koji slabo odražavaju stvarnu distribuciju, što može pogoršati točnost klasifikacije [12].

2.2.2. Postavke parametara korištene u literaturi

Kako je ranije navedeno, ključni parametri algoritma SMOTE uključuju broj najbližih susjeda **k** i broj sintetičkih uzoraka **q** koji se generiraju za svaki uzorak manjinske klase. Odabrane vrijednosti ovih parametara značajno utječu na karakteristike generiranih sintetičkih uzoraka i, posljedično, na učinkovitost klasifikatora. Ne postoji univerzalno optimalna kombinacija parametara, jer različite kombinacije mogu imati različite učinke na performanse ovisno o specifičnostima podataka i klasifikacijskih modela [13]. Tablica 2.2. prikazuje najčešće korištene postavke parametara algoritma SMOTE u analiziranim radovima iz relevantne literature [9, 11, 14-16].

Tablica 2.2. Korištene postavke parametara algoritma SMOTE u literaturi

Parametar	**Minimalna vrijednost u literaturi**	**Maksimalna vrijednost u literaturi**	**Najčešća vrijednost u literaturi**
k	3	5	5
q	1	30	1-5

U literaturi se za vrijednost parametra **k** gotovi isključivo koristi vrijednost 5 [9, 14-16]. Ova vrijednost je preporučena u izvornom SMOTE algoritmu zbog svoje umjerenosti i stabilnosti rezultata na različitim skupovima podataka. Ipak, korisno bi bilo istražiti kako upotreba manjih i većih vrijednosti od 5 utječe na performanse klasifikacijskih modela na skupovima s različitim stupnjevima neuravnoteženosti.

S druge strane, za parametar **q** u literaturi se najčešće koristi raspon vrijednosti od 1 do 5 [9, 11], ali u nekim radovima istraženi su i učinci vrijednosti **q** u znatno širem rasponu, kao što je slučaj u [15], gdje su eksperimenti provedeni i sa vrijednosti **q** do 30. Važno je i napomenuti da je u većini literature parametar **q** predstavljen kao postotak koliko će se izvorni skup podataka preuzorkovati, dok je u ovom radu prikazan kao cijeli broj. Stoga su vrijednosti prikazane u ovom radu pretvorene iz postotka u cijeli broj.

Performanse klasifikatora u navedenim istraživanjima nisu uvijek pokazivale jedinstven obrazac. Na nekim skupovima podataka povećanje vrijednosti parametra **q** poboljšalo je učinkovitost klasifikatora, dok je na drugima dovelo do značajnog pogoršanja. Stoga bi bilo važno proučiti kako ovaj parametar utječe na učinkovitost klasifikacijskih modela u različitim kontekstima podataka.

3. OSTVARENO PROGRAMSKO RJEŠENJE

Ostvareno programsko rješenje omogućuje učitavanje podataka u CSV (engl. *comma separared values*) formatu te provođenje klasifikacije tih podataka s više različitih klasifikacijskih modela. Osim same klasifikacije, postoje i mogućnosti poduzorkovanja i preuzorkovanja učitanog skupa podataka. Poduzorkovanje je omogućeno u vidu nasumičnog poduzorkovanja koji smanjuje neuravnoteženost skupa tako što iz skupa podataka uklanja nasumične uzorke većinske klase sve dok se broj uzoraka većinske i manjinske klase ne izjednači. S druge strane, preuzorkovanje je omogućeno putem SMOTE algoritma, koji generira sintetičke uzorke manjinske klase, pri čemu se mogu postaviti parametri **k** i **q** kako bi se prilagodio proces stvaranja tih uzoraka. Programsko rješenje razvijeno je u Python programskom jeziku koristeći biblioteke *pandas* i *numpy* za rad s podacima te biblioteke *tkinter* za izradu grafičkog korisničkog sučelja.

3.1. Način rada programskog rješenja

Razvijeno programsko rješenje implementirano je kao Python funkcija koja prihvaća sljedeće ulazne parametre: putanju do skupa podataka za eksperiment, broj ponavljanja eksperimenta, omjer podjele skupa podataka na podskupove za treniranje i testiranje, opciju standardizacije značajki skupa, izbor korištenja *baseline* klasifikacije bez promjena skupa podataka te korištenja klasifikacije s nasumično poduzorkovanim skupom radi usporedbe sa samostalno implementiranim SMOTE algoritmom, parametre SMOTE algoritma te odabir klasifikatora uz mogućnost podešavanja njegovih parametara.

Program započinje učitavanjem skupa podataka iz CSV datoteke i pohranjuje ih u *pandas DataFrame* objekt. Zatim se na učitanom skupu podataka vrši kodiranje pozitivne klase u vrijednost 1, a negativne u vrijednost 0. Nakon pripreme, eksperiment se provodi onoliko puta koliko je specificirano u ulaznim parametrima. U svakom ponavljanju vrši se podjela skupa na podskupove za treniranje i testiranje u zadanom omjeru te svako ponavljanje koristi novu početnu vrijednost (engl. *seed*) generatora pseudo-slučajnih brojeva koji se koristi kod podjele. Nakon ove podjele, na oba podskupa se vrši izdvajajanje stupaca u zasebne *DataFrame* objekte od kojih jedan sadrži značajke podataka, a drugi klasne oznake. Zatim, ukoliko je odabrana opcija za standardizaciju, primjenjuje se standardizacija pomoću *sklearn* klase *StandardScaler*. Standardizacija značajki osigurava da sve varijable imaju istu skalu, što može poboljšati performanse modela, posebno kod algoritama poput SVM-a i *k*-NN-a, koji su osjetljivi na udaljenosti među podacima.

Nakon toga, program započinje klasifikaciju nad skupom podataka koristeći odabrane klasifikatore. U trenutnoj implementaciji su dostupni klasifikatori: *k*-NN, GNB, SVM te stablo odluke. Prvo se klasificira bez promjene skupa podataka, ako je ta opcija odabrana. Zatim se, ukoliko je ta opcija odabrana, obavlja klasifikacija na podacima koji su nasumično poduzorkovani pomoću *sklearn* klase *RandomUnderSampler*. Na kraju, klasificiraju se podaci koji su preuzorkovani pomoću SMOTE algoritma s navedenim parametrima. Nakon svake klasifikacije, izračunavaju se F-1 i G-mean mjere, a rezultati se pohranjuju. F-1 mjera je od posebnog interesa kada se razmatraju učinkovitost klasifikacije manjinske klase, a G-Mean mjera je dobar pokazatelj uravnoteženosti performansi na obje klase. Na kraju, rezultati se formatiraju i izračunavaju prosječne vrijednosti mjera, koje se zatim prosljeđuju kao izlaz programa.

3.2. Prikaz i način oporabe programskog rješenja

Ostvareno programsko rješenje omogućava korisniku interakciju putem grafičkog korisničkog sučelja (engl. *graphical user interface*, GUI) koje pojednostavljuje proces izvođenja klasifikacijskih eksperimenata i prilagodbe parametara. Kroz GUI korisnik može lako upravljati postupkom pripreme podataka, podešavanja modela te izvođenja i analize rezultata. Korisniku su omogućene sljedeće funkcionalnosti preko GUI-a:

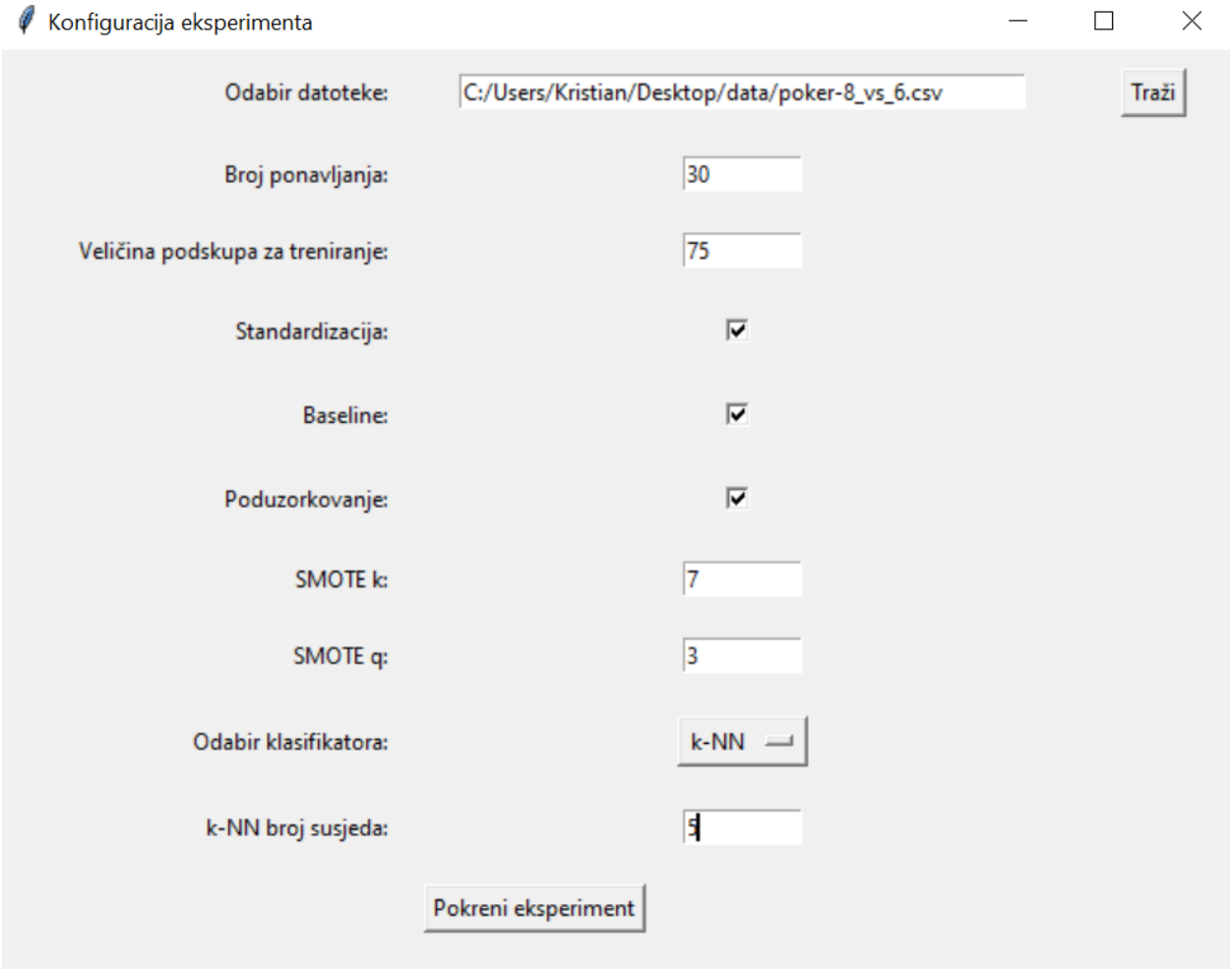
- **Učitavanje podataka iz CSV datoteke.** Korisnik može ručno unijeti putanju do datoteke u kojoj se nalazi skup podataka ili, putem integriranog dijaloga za pretraživanje datotečnog sustava, odabrati CSV datoteku s podacima. Ova funkcionalnost olakšava unos podataka i omogućuje brzu promjenu skupa podataka za različite eksperimente bez potrebe za ručnim uređivanjem koda.
- **Postavljanje broja ponavljanja izvođenja cijelog postupka.** Korisniku je omogućeno definiranje broja ponavljanja eksperimenta. Ova opcija je korisna za procjenu stabilnosti modela na istom skupu podataka kroz više ponavljanja, s različitim početnim vrijednostima za slučajne procese (npr. podjelu skupa podataka na podskupove za treniranje i testiranje). Na taj način korisnik može analizirati varijacije u rezultatima (F-1 mjera i G-mean mjera) kroz više ponavljanja.
- **Definiranje veličine podskupa za treniranje.** Korisnik može prilagoditi omjer podjele skupa podataka na podskupove za treniranje i testiranje. Time se omogućuje fleksibilnost u izvođenju eksperimenata jer se različiti omjeri mogu koristiti za optimizaciju performansi modela ili testiranje njegove robusnosti. Postavljanje manjeg podskupa za

treniranje može testirati sposobnost modela da uči iz malog uzorka, dok veći podskupovi za treniranje mogu pružiti preciznije rezultate.

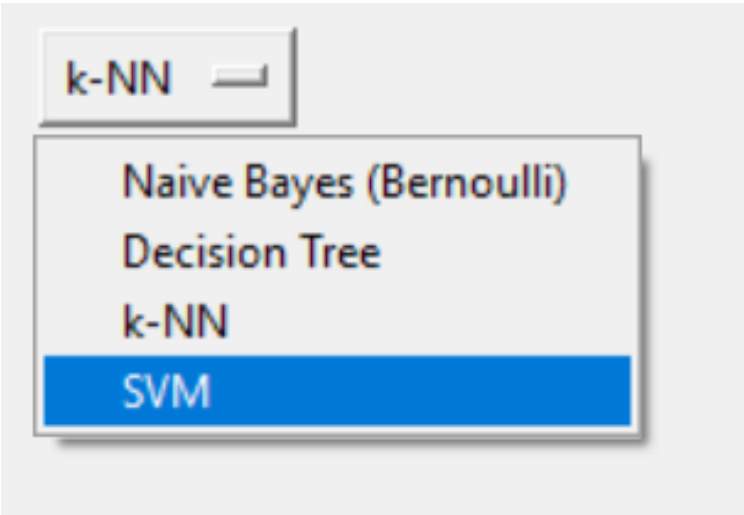
- **Odabir standardizacije ulaznih veličina.** Standardizacija ulaznih značajki važan je korak kada se radi s algoritmima osjetljivima na udaljenosti između podataka (npr. *k*-NN, SVM). Korisniku je omogućeno uključivanje ili isključivanje ove opcije, ovisno o tome želi li da značajke imaju normalizirane vrijednosti. Standardizacija pretvara značajke tako da imaju srednju vrijednost nula i standardnu devijaciju jedan.
- **Odabir hoće li se provesti *baseline* klasifikacija.** Korisnik može odabrati hoće li provesti *baseline* klasifikaciju, što podrazumijeva provođenje klasifikacije na izvornim podacima bez primjene bilo kakvih metoda uravnoteženja skupa podataka. Ova opcija omogućuje korisniku usporedbu performansi modela s drugim tehnikama koje se koriste za rješavanje problema neuravnoteženosti klasa.
- **Odabir hoće li se provesti klasifikacija s nasumičnim poduzorkovanjem.** Nasumično poduzorkovanje koristi se kako bi se smanjila dominacija većinske klase u skupu podataka. Korisnik može odlučiti želi li provesti klasifikaciju koristeći ovaj pristup te usporediti rezultate s *baseline* modelom i SMOTE-om.
- **Postavljanje parametara algoritma SMOTE.** Korisnik može postaviti broj najbližih susjeda manjinske klase (**k**) koji se koriste za generiranje novih primjera te broj uzoraka (**q**) koji će se generirati za svaki postojeći uzorak manjinske klase. Ova fleksibilnost omogućava testiranje različitih postavki parametara SMOTE algoritma kako bi se pronašla najbolja strategija uravnoteženja podataka.
- **Odabir klasifikatora.** Korisniku su na raspolaganju četiri klasifikatora: *k*-NN, GNB, SVM te stablo odluke. Mogućnost odabira različitih algoritama omogućuje usporedbu njihove učinkovitosti na istom skupu podataka. Svaki od tih klasifikatora ima svoje specifičnosti te će korisnik moći vidjeti koji bolje odgovara određenom problemu.
- **Postavljanje specifičnih parametara odabranog klasifikatora.** Ukoliko korisnik odabere *k*-NN kao klasifikator, može postaviti broj susjeda (*k*) koji će se koristiti u klasifikaciji. Time se omogućava postaljanje parametara algoritma, što može značajno utjecati na performanse modela. Sučelje pruža fleksibilnost da se različiti parametri lako testiraju.

Grafičko korisničko sučelje, prikazano na slici 3.1, korisniku omogućuje jednostavnu navigaciju i upravljanje eksperimentima. Na slici 3.2 prikazan je postupak odabira klasifikatora, gdje se jasno vidi kako korisnik može odabrati jedan ili više algoritama. Kroz ove funkcionalnosti,

korisniku je pružena potpuna kontrola nad konfiguracijom i izvođenjem eksperimenata, čime se olakšava analiza i optimizacija modela.

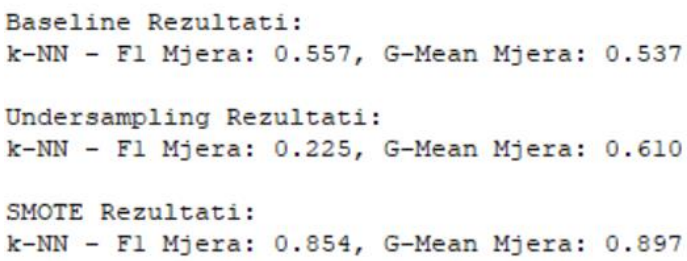


Slika 3.1. *Prikaz grafičkog korisničkog sučelja programskog rješenja*



Slika 3.2. *Prikaz odabira klasifikatora*

Nakon unosa podataka i odabira eksperimentalnih postavki, pritiskom na gumb 'Pokreni eksperiment' započinje se proces izvođenja eksperimenta. Ovaj gumb proslijeđuje postavljene opcije funkciji koja provodi eksperiment i vraća formatirane rezultate klasifikacije. Prikaz rezultata na korisničkom sučelju omogućuje korisniku da jednostavno usporedi učinak različitih klasifikatora i pristupa uravnoteženju podataka, te donese informirane odluke o najboljem pristupu za određeni skup podataka. Vizualni prikaz rezultata, prikazan na slici 3.3, sadrži prosječne vrijednosti F-1 i G-Mean mjera za sve odabrane pristupe klasifikaciji, olakšavajući interpretaciju rezultata i daljnju analizu.



Slika 3.3. *Prikaz rezultata eksperimenta na grafičkom korisničkom sučelju*

4. EKSPERIMENTALNA ANALIZA

Glavni cilj ovog rada je istražiti utjecaj različitih postavki parametara SMOTE algoritma na performanse klasifikacijskih modela kada se primjenjuju na neuravnotežene skupove podataka. U sklopu eksperimentalne analize, korišteno je sedam različitih skupova podataka, koji su zbog jednostavnijeg prikaza rezultata u daljnjem tekstu označeni oznakama D1 do D7. Ključne karakteristike tih skupova, kao što su broj uzoraka (redaka), broj značajki (stupaca) te omjer neuravnoteženosti broja uzoraka većinske i manjinske klase (engl. *imbalance ratio*, IR), prikazane su u Tablici 4.1.

Svi skupovi podataka korišteni u eksperimentu sadrže uzorke razvrstane u dvije klase: pozitivnu i negativnu. Ovi skupovi obuhvaćaju različite domene, uključujući detekciju bolesti (Pima, D1), klasifikaciju industrijskih procesa (Yeast3, D3) i kvalitativnu analizu vina (Winequality-red-4, D5). Unatoč različitim područjima primjene, svi skupovi dijele zajedničku karakteristiku neuravnoteženosti klasa, gdje je jedna klasa značajno brojnija od druge. Važno je napomenuti da prva tri skupa podataka imaju omjer neuravnoteženosti (IR) manji od 10, dok preostala četiri skupa imaju IR veći od 10. Ova raznolikost u omjerima neuravnoteženosti omogućuje istraživanje SMOTE algoritma u različitim stupnjevima neuravnoteženosti, što je ključno za razumijevanje njegovog utjecaja na performanse klasifikacijskih modela.

Eksperiment je izvršen na odabranim skupovima podataka u tri različita scenarija: bez izmjene podataka (engl. *baseline*), uz primjenu nasumičnog poduzorkovanja, te uz korištenje SMOTE algoritma s različitim postavkama parametara. Za parametar **k** testirane su vrijednosti 3, 5, 7 i 10, dok su za parametar **q** korištene vrijednosti 1, 3, 5, 10, 15, 20 i 30. Ove vrijednosti odabrane su na temelju literature te prethodnih eksperimenata.

Tablica 4.1. Korišteni skupovi podataka i njihove karakteristike

Naziv skupa podataka	**Oznaka**	**Broj uzoraka**	**Broj značajki**	**IR**
Pima	D1	768	8	1.87
Vehicle1	D2	846	18	2.9
Yeast3	D3	1484	8	8.1
Shuttle-c0-vs-c4	D4	1829	9	13.87
Winequality-red-4	D5	1599	11	29.17
Poker-8_vs_6	D6	1477	10	85.88
Abalone19	D7	4174	8	129.44

4.1. **Postavke eksperimenta**

Preuzeti skupovi podataka prvo su pretvoreni iz izvornog DAT formata (sirovi podaci) u CSV format (podaci odvojeni zarezima) radi lakše obrade i analize. Svaki skup podataka zatim je podijeljen na podskupove za treniranje i testiranje u omjeru 75:25, pri čemu je omjer neuravnoteženosti (IR) očuvan u oba podskupa.

Eksperimentalna analiza provedena je s tri različita pristupa: SMOTE algoritam s različitim kombinacijama parametara, *baseline* (bez preuzorkovanja), te nasumično poduzorkovanje. Svaki od ovih pristupa testiran je s podijeljenim podacima kroz 30 ponavljanja. Pri svakom ponavljanju, podaci su ponovno podijeljeni na skup za treniranje i testiranje kako bi se osigurala robusnost rezultata.

Za klasifikaciju su korišteni sljedeći modeli: 5-NN, stablo odluke, GNB, te SVM s radijalnom funkcijom jezgre. Važno je napomenuti da su za svako ponavljanje svi klasifikatori koristili iste podskupove za treniranje i testiranje, što je omogućilo izravnu usporedbu njihovih performansi.

Prije podjele podataka na skupove za treniranje i testiranje, podaci su standardizirani tako da sve značajke imaju srednju vrijednost 0 i standardnu devijaciju 1, kako bi se osigurala ujednačenost značajki. Za skup podataka Abalone19, kategorijska značajka spol kodirana je u numeričke vrijednosti -1, 0 i 1. Nije bilo potrebe za uklanjanjem nedostajućih ili dupliciranih vrijednosti, budući da takvi podaci nisu bili prisutni.

Tijekom izvođenja algoritama prikupljene su mjere F-1 i G-mean, koje su prikazane u rezultatima u obliku srednje vrijednosti na 30 ponavljanja ± standardna devijacija. Ove statističke mjere omogućuju detaljnu analizu performansi klasifikatora pri radu s neuravnoteženim podacima.

4.2. **Rezultati**

Rezultati eksperimenta za klasifikacijski model 5-NN prikazani su u tablicama 4.2 i 4.3. Tablica 4.2 prikazuje prosječne vrijednosti F-1 mjere na 30 ponavljanja, dok tablica 4.3 prikazuje prosječne vrijednosti G-mean mjere, uz pripadajuće standardne devijacije. Najviše prosječne vrijednosti za svaki skup podataka prikazane su podebljano.

Rezultati za *baseline* pristup pokazuju stabilne, ali ne izvanredne performanse, s umjerenim F-1 vrijednostima, osim na skupovima D3 i D4 gdje su one izrazito visoke. G-mean mjere su nešto niže, ali pokazuju sličan trend, reflektirajući sklonost modela prema većinskoj klasi.

Primjena nasumičnog poduzorkovanja daje mješovite rezultate. Na nekim skupovima, poput D1, dolazi do blagog poboljšanja F-1 mjere u odnosu na *baseline,* dok na skupovima s visokim omjerom neuravnoteženosti, kao što su D6 i D7, dolazi do značajnog pada performansi. Unatoč tome, poduzorkovanje općenito poboljšava G-mean vrijednosti, posebno na skupovima kao što su D5 i D7, sugerirajući da ova metoda bolje usklađuje osjetljivost i specifičnost modela.

SMOTE algoritam, s različitim postavkama parametara, pokazuje koliko su ti parametri ključni za performanse modela. Manje vrijednosti parametra **q** (npr. **q**=1-3) često poboljšavaju F-1 mjeru, osobito na skupovima s velikom neuravnoteženošću, poput D5 i D6. Međutim, povećanje **q** može dovesti do pada F-1 mjere zbog prekomjernog generiranja uzoraka. S druge strane, G-mean mjera se generalno poboljšava s povećanjem **q**, posebno na skupovima s visokom neuravnoteženošću.

Ovi rezultati pokazuju da različite tehnike poduzorkovanja i preuzorkovanja imaju različit utjecaj na učinkovitost 5-NN klasifikatora. Zbog prirode algoritma koji se oslanja na neposredne susjede, neravnoteža u klasama može dovesti do pristranosti prema većinskoj klasi, što dodatno naglašava važnost pažljive primjene tehnika uravnoteženja podataka. Dok *baseline* pristup nudi solidne, ali ne i izvanredne rezultate, poduzorkovanje poboljšava G-mean mjeru uz čest pad F-1 mjere, što ukazuje na kompromis između osjetljivosti i specifičnosti modela. SMOTE metoda ima najveći potencijal za poboljšanje performansi, posebno uz pažljivo podešavanje parametara, iako postoji rizik od smanjenja F-1 mjere zbog prekomjernog generiranja sintetičkih podataka. Pretjerano generiranje sintetičkih uzoraka može stvoriti podatke koji značajno odstupaju od stvarne razdiobe izvornog skupa podataka, što smanjuje učinkovitost modela. Stoga, prilikom primjene SMOTE-a, važno je održati ravnotežu između povećanja osjetljivosti modela i održavanja njegove sposobnosti generalizacije, kako bi se izbjegla prenaučenost i zadržala konzistentna preciznost modela na različitim skupovima podataka.

Tablica 4.2. Rezultati F-1 mjere za klasifikacijski model 5-NN

Algoritam	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Baseline	0.694±0.03	0.694±0.03	**0.855±0.03**	**0.996±0.00**	0.507±0.03	0.598±0.10	0.498±0.00
Poduzorkovanje	**0.698±0.03**	0.681±0.02	0.794±0.02	0.995±0.01	0.477±0.03	0.249±0.05	0.381±0.03
SMOTE(3,1)	0.689±0.03	0.710±0.03	0.835±0.02	0.996±0.01	0.547±0.04	0.786±0.09	0.500±0.02
SMOTE(3,3)	0.592±0.03	0.661±0.02	0.740±0.02	0.996±0.01	0.549±0.04	0.837±0.10	0.508±0.02
SMOTE(3,5)	0.531±0.04	0.621±0.02	0.673±0.02	0.990±0.01	0.542±0.03	0.836±0.06	0.510±0.02
SMOTE(3,10)	0.434±0.04	0.563±0.03	0.574±0.02	0.953±0.02	0.520±0.02	0.781±0.07	0.503±0.02
SMOTE(3,15)	0.397±0.03	0.531±0.03	0.506±0.02	0.837±0.03	0.505±0.02	0.753±0.06	0.503±0.01
SMOTE(3,20)	0.373±0.03	0.509±0.04	0.457±0.02	0.762±0.03	0.494±0.02	0.729±0.05	0.502±0.01
SMOTE(3,30)	0.356±0.03	0.480±0.04	0.380±0.02	0.498±0.02	0.476±0.02	0.706±0.05	0.502±0.01
SMOTE(5,1)	0.688±0.03	**0.713±0.02**	0.832±0.03	0.996±0.01	**0.558±0.05**	0.803±0.12	0.500±0.02
SMOTE(5,3)	0.587±0.03	0.657±0.02	0.739±0.02	0.996±0.01	0.556±0.04	0.866±0.09	**0.512±0.02**
SMOTE(5,5)	0.519±0.04	0.614±0.03	0.672±0.02	0.990±0.01	0.546±0.04	0.844±0.07	0.507±0.02
SMOTE(5,10)	0.434±0.03	0.553±0.03	0.568±0.02	0.954±0.02	0.525±0.03	0.799±0.08	0.511±0.02
SMOTE(5,15)	0.385±0.03	0.517±0.03	0.498±0.02	0.837±0.03	0.505±0.02	0.753±0.06	0.504±0.02
SMOTE(5,20)	0.365±0.02	0.489±0.03	0.447±0.02	0.760±0.03	0.494±0.02	0.728±0.05	0.504±0.01
SMOTE(5,30)	0.341±0.02	0.461±0.04	0.368±0.02	0.497±0.02	0.470±0.02	0.680±0.03	0.500±0.01
SMOTE(7,1)	0.688±0.03	0.706±0.03	0.835±0.03	0.996±0.01	0.557±0.05	0.776±0.10	0.503±0.03
SMOTE(7,3)	0.592±0.03	0.657±0.03	0.736±0.02	0.996±0.01	0.553±0.04	**0.867±0.08**	0.503±0.02
SMOTE(7,5)	0.520±0.04	0.615±0.02	0.671±0.02	0.990±0.01	0.547±0.04	0.845±0.09	0.512±0.03
SMOTE(7,10)	0.422±0.03	0.550±0.03	0.568±0.02	0.952±0.02	0.528±0.03	0.789±0.07	0.509±0.02
SMOTE(7,15)	0.378±0.03	0.510±0.03	0.493±0.02	0.834±0.03	0.509±0.02	0.741±0.06	0.506±0.02
SMOTE(7,20)	0.355±0.03	0.482±0.04	0.442±0.02	0.760±0.03	0.493±0.02	0.713±0.05	0.505±0.01
SMOTE(7,30)	0.332±0.03	0.449±0.04	0.362±0.02	0.495±0.03	0.470±0.02	0.667±0.04	0.499±0.01
SMOTE(10,1)	0.687±0.03	0.704±0.02	0.837±0.02	0.996±0.01	0.550±0.04	0.826±0.09	0.503±0.02
SMOTE(10,3)	0.594±0.03	0.655±0.02	0.739±0.02	0.996±0.01	0.555±0.04	0.864±0.08	0.509±0.04
SMOTE(10,5)	0.506±0.04	0.611±0.03	0.671±0.02	0.990±0.01	0.544±0.03	0.844±0.08	0.512±0.03
SMOTE(10,10)	0.410±0.03	0.547±0.03	0.564±0.02	0.953±0.02	0.523±0.02	0.779±0.07	0.511±0.02
SMOTE(10,15)	0.374±0.03	0.504±0.03	0.488±0.02	0.836±0.03	0.511±0.03	0.737±0.06	0.509±0.02
SMOTE(10,20)	0.350±0.03	0.474±0.03	0.436±0.02	0.759±0.03	0.494±0.02	0.705±0.04	0.507±0.02
SMOTE(10,30)	0.323±0.02	0.436±0.04	0.357±0.02	0.493±0.03	0.470±0.02	0.664±0.03	0.497±0.01

Tablica 4.3. Rezultati G-mean mjere za klasifikacijski model 5-NN

Algoritam	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Baseline	0.688±0.03	0.691±0.04	0.835±0.03	0.992±0.01	0.508±0.02	0.562±0.06	0.500±0.00
Poduzorkovanje	0.710±0.03	0.744±0.03	**0.902±0.02**	**0.995±0.01**	**0.679±0.06**	0.610±0.09	**0.681±0.08**
SMOTE(3,1)	**0.718±0.03**	0.759±0.03	0.886±0.03	0.992±0.01	0.540±0.04	0.725±0.10	0.501±0.01
SMOTE(3,3)	0.666±0.02	0.751±0.02	0.875±0.02	0.992±0.01	0.575±0.06	0.864±0.10	0.512±0.03
SMOTE(3,5)	0.630±0.03	0.730±0.02	0.848±0.02	0.991±0.01	0.593±0.06	0.917±0.08	0.524±0.05
SMOTE(3,10)	0.575±0.02	0.694±0.02	0.793±0.02	0.986±0.01	0.602±0.05	0.947±0.07	0.526±0.05
SMOTE(3,15)	0.557±0.02	0.675±0.03	0.746±0.02	0.962±0.01	0.603±0.06	0.965±0.06	0.545±0.05
SMOTE(3,20)	0.546±0.02	0.663±0.03	0.711±0.02	0.940±0.02	0.611±0.06	0.963±0.06	0.555±0.05
SMOTE(3,30)	0.539±0.02	0.646±0.03	0.661±0.02	0.786±0.02	0.610±0.07	0.973±0.05	0.598±0.05
SMOTE(5,1)	0.716±0.03	**0.761±0.03**	0.883±0.03	0.992±0.01	0.550±0.05	0.750±0.11	0.501±0.01
SMOTE(5,3)	0.661±0.02	0.748±0.03	0.873±0.02	0.992±0.01	0.585±0.06	0.898±0.08	0.517±0.04
SMOTE(5,5)	0.623±0.03	0.729±0.03	0.845±0.02	0.991±0.01	0.599±0.07	0.938±0.08	0.518±0.04
SMOTE(5,10)	0.576±0.02	0.689±0.02	0.789±0.02	0.986±0.01	0.617±0.07	0.964±0.06	0.548±0.05
SMOTE(5,15)	0.552±0.02	0.669±0.02	0.741±0.02	0.962±0.01	0.613±0.06	0.977±0.04	0.552±0.06
SMOTE(5,20)	0.543±0.01	0.652±0.02	0.704±0.02	0.939±0.02	0.619±0.06	0.975±0.04	0.575±0.06
SMOTE(5,30)	0.533±0.01	0.636±0.02	0.653±0.02	0.784±0.02	0.607±0.07	0.973±0.03	0.612±0.06
SMOTE(7,1)	0.715±0.03	0.754±0.04	0.889±0.03	0.992±0.01	0.549±0.05	0.716±0.10	0.503±0.02
SMOTE(7,3)	0.667±0.03	0.748±0.03	0.872±0.02	0.992±0.01	0.581±0.06	0.902±0.09	0.504±0.02
SMOTE(7,5)	0.624±0.03	0.730±0.02	0.847±0.02	0.991±0.01	0.601±0.06	0.925±0.08	0.525±0.05
SMOTE(7,10)	0.571±0.02	0.689±0.02	0.790±0.02	0.986±0.01	0.623±0.06	0.967±0.06	0.543±0.05
SMOTE(7,15)	0.549±0.02	0.665±0.02	0.737±0.02	0.962±0.01	0.622±0.06	0.972±0.06	0.562±0.06
SMOTE(7,20)	0.540±0.02	0.648±0.02	0.700±0.02	0.939±0.02	0.614±0.06	0.977±0.03	0.590±0.06
SMOTE(7,30)	0.530±0.01	0.628±0.03	0.649±0.02	0.783±0.02	0.611±0.07	0.971±0.05	0.622±0.06
SMOTE(10,1)	0.716±0.03	0.751±0.03	0.888±0.03	0.992±0.01	0.542±0.03	0.766±0.10	0.503±0.02
SMOTE(10,3)	0.670±0.03	0.749±0.02	0.875±0.02	0.992±0.01	0.582±0.05	0.881±0.09	0.510±0.04
SMOTE(10,5)	0.615±0.03	0.728±0.02	0.846±0.02	0.991±0.01	0.595±0.06	0.921±0.09	0.522±0.04
SMOTE(10,10)	0.565±0.02	0.689±0.02	0.788±0.02	0.986±0.01	0.613±0.06	0.959±0.06	0.548±0.06
SMOTE(10,15)	0.549±0.02	0.662±0.02	0.734±0.02	0.962±0.01	0.628±0.06	0.971±0.04	0.575±0.07
SMOTE(10,20)	0.538±0.01	0.643±0.02	0.697±0.02	0.939±0.02	0.620±0.07	**0.981±0.02**	0.598±0.07
SMOTE(10,30)	0.526±0.01	0.621±0.02	0.646±0.01	0.781±0.02	0.608±0.07	0.975±0.02	0.616±0.07

Rezultati eksperimenta za klasifikacijski model stabla odluke prikazani su u Tablicama 4.4 i 4.5. Tablica 4.4 prikazuje prosječne vrijednosti F-1 mjere na 30 ponavljanja, dok Tablica 4.5 prikazuje prosječne vrijednosti G-mean mjere, uz pripadajuće standardne devijacije. Najviše prosječne vrijednosti za svaki skup podataka prikazane su podebljano.

Rezultati za *baseline* pristup pokazuju solidne performanse, s najvišim F-1 vrijednostima na skupovima D1, D3, D4, D5 i D7. Posebno je zanimljivo primijetiti da *baseline* pristup na skupu D4 postiže F-1 i G-mean rezultat od 1.0, što ukazuje na savršenu klasifikaciju u tom slučaju. Ovakav rezultat sugerira da promjena skupa podataka pri korištenju stabla odluke ponekad nema smisla, jer je poboljšanje performansi nemoguće u situacijama u kojima je već postignut maksimalan rezultat.

Primjena nasumičnog poduzorkovanja ponovno donosi mješovite rezultate. Na primjer, na skupu D2 dolazi do blagog poboljšanja F-1 mjere u odnosu na *baseline,* dok na skupovima s velikom neuravnoteženošću, kao što su D5, D6 i D7, dolazi do značajnog pada performansi. S druge strane, poduzorkovanje generalno poboljšava G-mean vrijednosti, posebno na skupovima D2, D3, D5 i D7, gdje često daje najbolje rezultate u pogledu G-mean mjere.

Primjena SMOTE algoritma pokazuje povećanje vrijednosti F-1 mjere u odnosu na *baseline* samo u skupovima D2 i D6, dok u ostalim skupovima F-1 mjera opada, posebno uz povećanje parametra **q**, što dovodi do značajnog smanjenja performansi. S druge strane, G-mean mjera se općenito povećava primjenom SMOTE algoritma. U skupovima s nižim omjerom neuravnoteženosti, to povećanje je blago, dok je kod skupova s većim omjerom neuravnoteženosti, posebno D6, povećanje G-mean mjere značajno. Na primjer, na skupu D6, najviša G-mean vrijednost zabilježena je za najveću vrijednost parametra **q** (u ovom slučaju **q**=30), dok na ostalim skupovima povećanje parametra **q** uglavnom dovodi do pada G-mean mjere.

U konačnici, ovi rezultati pokazuju da stabla odluke često ne profitiraju od preuzorkovanja, a primjena SMOTE-a može imati neznatan ili čak negativan učinak na performanse, ovisno o skupu podataka. Ipak, u nekim slučajevima, kao što je skup D6, SMOTE može biti koristan i dovesti do značajnih poboljšanja performansi stabla odluke.

Tablica 4.4. Rezultati F-1 mjere za klasifikacijski model stablo odluke

Algoritam	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Baseline	**0.677±0.04**	0.655±0.04	**0.815±0.03**	**1.000±0.00**	**0.537±0.04**	0.592±0.17	**0.519±0.03**
Poduzorkovanje	0.665±0.04	0.664±0.03	0.789±0.03	0.999±0.00	0.432±0.03	0.381±0.07	0.418±0.03
SMOTE(3,1)	0.660±0.03	0.667±0.03	0.768±0.04	0.991±0.01	0.513±0.03	0.651±0.19	0.517±0.03
SMOTE(3,3)	0.606±0.04	0.638±0.04	0.695±0.04	0.968±0.02	0.531±0.03	0.613±0.19	0.500±0.02
SMOTE(3,5)	0.552±0.04	0.632±0.04	0.650±0.04	0.854±0.05	0.528±0.04	0.473±0.07	0.496±0.02
SMOTE(3,10)	0.496±0.04	0.590±0.04	0.549±0.03	0.712±0.04	0.512±0.02	0.457±0.04	0.494±0.02
SMOTE(3,15)	0.467±0.04	0.566±0.03	0.479±0.04	0.638±0.03	0.494±0.03	0.461±0.05	0.497±0.03
SMOTE(3,20)	0.461±0.03	0.555±0.04	0.428±0.04	0.588±0.02	0.499±0.02	0.467±0.06	0.493±0.03
SMOTE(3,30)	0.443±0.03	0.564±0.04	0.376±0.02	0.518±0.02	0.483±0.02	0.487±0.06	0.496±0.03
SMOTE(5,1)	0.655±0.04	0.670±0.03	0.781±0.04	0.994±0.01	0.530±0.04	0.622±0.19	0.517±0.03
SMOTE(5,3)	0.604±0.03	0.641±0.02	0.692±0.03	0.970±0.02	0.519±0.03	0.635±0.20	0.501±0.02
SMOTE(5,5)	0.544±0.03	0.632±0.04	0.632±0.04	0.865±0.05	0.523±0.02	0.515±0.13	0.493±0.02
SMOTE(5,10)	0.491±0.04	0.573±0.03	0.541±0.03	0.716±0.04	0.500±0.03	0.431±0.02	0.499±0.02
SMOTE(5,15)	0.447±0.03	0.555±0.03	0.475±0.04	0.641±0.03	0.492±0.02	0.464±0.04	0.493±0.02
SMOTE(5,20)	0.434±0.04	0.555±0.04	0.419±0.04	0.591±0.02	0.486±0.02	0.485±0.05	0.488±0.03
SMOTE(5,30)	0.419±0.04	0.545±0.04	0.376±0.03	0.520±0.02	0.475±0.03	0.469±0.06	0.492±0.02
SMOTE(7,1)	0.659±0.04	**0.674±0.04**	0.782±0.04	0.992±0.01	0.532±0.05	**0.697±0.23**	0.509±0.02
SMOTE(7,3)	0.606±0.04	0.655±0.03	0.701±0.04	0.970±0.02	0.523±0.03	0.568±0.17	0.496±0.03
SMOTE(7,5)	0.535±0.03	0.625±0.03	0.631±0.04	0.869±0.05	0.521±0.03	0.473±0.07	0.498±0.03
SMOTE(7,10)	0.475±0.04	0.578±0.04	0.528±0.04	0.712±0.04	0.499±0.03	0.436±0.03	0.486±0.02
SMOTE(7,15)	0.448±0.03	0.552±0.03	0.463±0.04	0.639±0.03	0.498±0.03	0.462±0.05	0.494±0.02
SMOTE(7,20)	0.432±0.03	0.555±0.03	0.417±0.05	0.591±0.02	0.479±0.02	0.465±0.07	0.492±0.02
SMOTE(7,30)	0.408±0.03	0.544±0.04	0.372±0.04	0.520±0.02	0.473±0.02	0.480±0.07	0.484±0.02
SMOTE(10,1)	0.658±0.03	0.667±0.03	0.776±0.04	0.994±0.01	0.531±0.03	0.654±0.20	0.510±0.02
SMOTE(10,3)	0.599±0.05	0.644±0.03	0.707±0.04	0.965±0.02	0.526±0.03	0.577±0.17	0.501±0.02
SMOTE(10,5)	0.543±0.04	0.619±0.03	0.641±0.05	0.853±0.05	0.513±0.03	0.463±0.03	0.500±0.02
SMOTE(10,10)	0.474±0.04	0.566±0.04	0.536±0.03	0.711±0.04	0.505±0.03	0.435±0.02	0.490±0.02
SMOTE(10,15)	0.434±0.03	0.558±0.02	0.466±0.04	0.646±0.03	0.490±0.02	0.455±0.05	0.488±0.02
SMOTE(10,20)	0.422±0.03	0.544±0.03	0.409±0.04	0.589±0.02	0.481±0.02	0.460±0.06	0.484±0.02
SMOTE(10,30)	0.399±0.04	0.542±0.04	0.356±0.03	0.517±0.02	0.465±0.02	0.447±0.09	0.478±0.03

Tablica 4.5. Rezultati G-mean mjere za klasifikacijski model stablo odluke

Algoritam	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Baseline	0.677±0.04	0.658±0.04	0.818±0.04	**1.000±0.00**	0.540±0.05	0.591±0.17	0.523±0.04
Poduzorkovanje	**0.678±0.04**	**0.701±0.04**	**0.885±0.03**	**1.000±0.00**	**0.609±0.07**	0.632±0.14	**0.674±0.10**
SMOTE(3,1)	0.674±0.03	0.686±0.03	0.841±0.04	0.999±0.00	0.525±0.05	0.669±0.20	0.533±0.05
SMOTE(3,3)	0.645±0.03	0.670±0.04	0.836±0.04	0.995±0.00	0.571±0.06	0.698±0.17	0.544±0.06
SMOTE(3,5)	0.614±0.03	0.671±0.04	0.819±0.04	0.973±0.01	0.577±0.07	0.700±0.17	0.562±0.06
SMOTE(3,10)	0.586±0.03	0.640±0.03	0.757±0.03	0.926±0.02	0.571±0.05	0.766±0.15	0.583±0.07
SMOTE(3,15)	0.568±0.03	0.612±0.04	0.706±0.03	0.890±0.02	0.546±0.07	0.774±0.16	0.587±0.06
SMOTE(3,20)	0.568±0.03	0.610±0.04	0.670±0.04	0.858±0.01	0.559±0.07	0.776±0.16	0.572±0.08
SMOTE(3,30)	0.561±0.02	0.613±0.05	0.636±0.02	0.807±0.02	0.539±0.05	**0.809±0.17**	0.566±0.07
SMOTE(5,1)	0.669±0.04	0.690±0.03	0.843±0.03	0.999±0.00	0.553±0.06	0.633±0.19	0.529±0.05
SMOTE(5,3)	0.645±0.03	0.678±0.03	0.832±0.03	0.996±0.00	0.551±0.06	0.750±0.19	0.539±0.06
SMOTE(5,5)	0.611±0.03	0.672±0.04	0.811±0.03	0.975±0.01	0.577±0.05	0.689±0.18	0.551±0.07
SMOTE(5,10)	0.583±0.03	0.626±0.03	0.752±0.04	0.928±0.02	0.549±0.07	0.727±0.15	0.593±0.08
SMOTE(5,15)	0.562±0.03	0.609±0.04	0.701±0.04	0.892±0.02	0.555±0.06	0.761±0.16	0.594±0.08
SMOTE(5,20)	0.555±0.03	0.606±0.04	0.664±0.04	0.860±0.01	0.551±0.07	0.763±0.19	0.587±0.09
SMOTE(5,30)	0.550±0.03	0.596±0.04	0.639±0.03	0.809±0.01	0.541±0.07	0.763±0.19	0.584±0.08
SMOTE(7,1)	0.674±0.04	0.697±0.04	0.848±0.04	0.999±0.00	0.556±0.07	0.711±0.23	0.521±0.04
SMOTE(7,3)	0.648±0.04	0.691±0.04	0.840±0.04	0.996±0.00	0.560±0.05	0.688±0.18	0.549±0.07
SMOTE(7,5)	0.602±0.03	0.669±0.04	0.810±0.04	0.976±0.01	0.573±0.07	0.701±0.16	0.586±0.08
SMOTE(7,10)	0.573±0.03	0.632±0.04	0.742±0.04	0.927±0.01	0.548±0.05	0.726±0.15	0.580±0.07
SMOTE(7,15)	0.563±0.02	0.612±0.04	0.700±0.04	0.890±0.02	0.574±0.07	0.750±0.19	0.607±0.08
SMOTE(7,20)	0.557±0.02	0.611±0.04	0.665±0.04	0.861±0.01	0.547±0.06	0.759±0.17	0.608±0.06
SMOTE(7,30)	0.547±0.02	0.598±0.04	0.635±0.03	0.809±0.02	0.539±0.06	0.802±0.17	0.610±0.10
SMOTE(10,1)	0.672±0.03	0.690±0.04	0.846±0.03	0.999±0.00	0.553±0.05	0.651±0.19	0.520±0.03
SMOTE(10,3)	0.643±0.04	0.679±0.03	0.846±0.04	0.995±0.00	0.569±0.07	0.713±0.17	0.561±0.05
SMOTE(10,5)	0.611±0.04	0.664±0.04	0.809±0.04	0.973±0.01	0.562±0.06	0.637±0.16	0.595±0.08
SMOTE(10,10)	0.576±0.01	0.626±0.04	0.752±0.03	0.926±0.02	0.571±0.07	0.724±0.15	0.584±0.08
SMOTE(10,15)	0.556±0.03	0.617±0.03	0.700±0.03	0.894±0.02	0.551±0.05	0.728±0.18	0.618±0.08
SMOTE(10,20)	0.549±0.03	0.601±0.03	0.657±0.03	0.860±0.01	0.543±0.07	0.742±0.20	0.607±0.06
SMOTE(10,30)	0.544±0.03	0.603±0.04	0.629±0.03	0.807±0.02	0.533±0.06	0.765±0.18	0.616±0.09

Rezultati eksperimenta za klasifikacijski model GNB prikazani su u Tablicama 4.6 i 4.7. Tablica 4.6 prikazuje prosječne vrijednosti F-1 mjere na 30 ponavljanja, dok Tablica 4.7 prikazuje prosječne vrijednosti G-mean mjere, uz pripadajuće standardne devijacije. Najviše prosječne vrijednosti za svaki skup podataka prikazane su podebljano.

Baseline rezultati su stabilni, ali ne previsoki. Iako na nekoliko skupova ovaj pristup pruža najveće vrijednosti F-1 mjere, G-mean vrijednosti za te skupove su ipak niže u usporedbi s drugim pristupima. Također, značajno je napomenuti da je kod ovog klasifikatora najveći učinak SMOTE algoritma na skup podataka D4, koji je kod drugih klasifikatora gotovo optimalan na *baseline* pristupu, s vrijednostima vrlo blizu 1. Ipak, kod GNB klasifikatora, *baseline* rezultati na ovom skupu su nešto niži, ali su optimizirani pomoću SMOTE algoritma s niskim vrijednostima parametra **q**.

Poduzorkovanje daje slične rezultate kao i na drugim klasifikacijskim modelima s najčešće nižim vrijednostima F-1 mjere, ali višim vrijednostima G-mean mjere. Na primjer, kod skupa D3 može se zamijetiti da su rezultati poduzorkovanja znatno veći od *baseline* rezultata, ali primjena SMOTE algoritma i na tom skupu dovela je do još većih vrijednosti.

SMOTE algoritam dao je prilično dobre rezultate na ovom klasifikatoru, omogućivši postizanje najboljeg omjera vrijednosti F-1 i G-mean mjera. Kao što je prethodno spomenuto, skup D4 je gotovo savršeno klasificiran nakon primjene SMOTE-a s manjim vrijednostima parametra **q** (**q**=3 ili **q**=1). Međutim, na skupu D6 uočava se da je pozitivni učinak SMOTE-a znatno manji nego kod drugih klasifikatora. Vrijednosti parametra **k** ovisile su o pojedinačnim skupovima podataka; tako su na nekim najbolje performanse dosegnute s **k**=10, a na drugima s **k**=3. Parametar **q** ponovno je ovisio o omjeru neuravnoteženosti pojedinih skupova, ali je primjećeno opadanje performansi na nekim skupovima s visokim omjerom neuravnoteženosti kod velikih vrijednosti parametra **q** zbog prekomjernog generiranja sintetičkih uzoraka.

GNB klasifikator pokazao je mogućnost pozitivnog učinka SMOTE algoritma s dobro određenim postavkama parametara.

Tablica 4.6. Rezultati F-1 mjere za klasifikacijski model GNB

Algoritam	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Baseline	**0.683 ± 0.03**	**0.612 ± 0.03**	0.471 ± 0.00	0.958 ± 0.03	0.492 ± 0.00	0.497 ± 0.00	**0.498 ± 0.00**
Poduzorkovanje	0.680 ± 0.03	0.608 ± 0.03	0.627 ± 0.03	0.995 ± 0.01	0.414 ± 0.02	0.342 ± 0.04	0.362 ± 0.01
SMOTE(3, 1)	0.673 ± 0.03	0.605 ± 0.03	0.710 ± 0.04	**0.997 ± 0.00**	0.497 ± 0.02	0.497 ± 0.00	0.495 ± 0.01
SMOTE(3, 3)	0.617 ± 0.03	0.597 ± 0.03	0.619 ± 0.03	**0.997 ± 0.00**	0.541 ± 0.03	0.497 ± 0.00	0.448 ± 0.04
SMOTE(3, 5)	0.553 ± 0.03	0.588 ± 0.03	0.593 ± 0.03	0.867 ± 0.08	0.528 ± 0.03	0.497 ± 0.00	0.407 ± 0.03
SMOTE(3, 10)	0.454 ± 0.03	0.569 ± 0.02	0.587 ± 0.03	0.628 ± 0.02	0.475 ± 0.03	0.511 ± 0.05	0.388 ± 0.01
SMOTE(3, 15)	0.369 ± 0.03	0.553 ± 0.03	0.578 ± 0.03	0.613 ± 0.03	0.442 ± 0.02	0.543 ± 0.08	0.386 ± 0.01
SMOTE(3, 20)	0.329 ± 0.02	0.540 ± 0.03	0.486 ± 0.03	0.603 ± 0.02	0.418 ± 0.02	0.541 ± 0.07	0.383 ± 0.01
SMOTE(3, 30)	0.268 ± 0.03	0.524 ± 0.03	0.326 ± 0.03	0.525 ± 0.05	0.379 ± 0.02	**0.550 ± 0.07**	0.381 ± 0.01
SMOTE(5, 1)	0.669 ± 0.03	0.605 ± 0.03	0.710 ± 0.04	**0.997 ± 0.00**	0.497 ± 0.02	0.497 ± 0.00	0.494 ± 0.01
SMOTE(5, 3)	0.617 ± 0.03	0.596 ± 0.03	0.621 ± 0.03	**0.997 ± 0.00**	**0.542 ± 0.03**	0.497 ± 0.00	0.448 ± 0.04
SMOTE(5, 5)	0.556 ± 0.03	0.587 ± 0.03	0.594 ± 0.03	0.874 ± 0.08	0.529 ± 0.03	0.497 ± 0.00	0.410 ± 0.03
SMOTE(5, 10)	0.455 ± 0.03	0.569 ± 0.03	0.587 ± 0.03	0.628 ± 0.02	0.477 ± 0.03	0.497 ± 0.00	0.388 ± 0.01
SMOTE(5, 15)	0.370 ± 0.03	0.553 ± 0.03	0.579 ± 0.03	0.614 ± 0.02	0.442 ± 0.02	0.522 ± 0.07	0.385 ± 0.01
SMOTE(5, 20)	0.331 ± 0.02	0.539 ± 0.03	0.485 ± 0.03	0.606 ± 0.02	0.417 ± 0.02	0.537 ± 0.07	0.383 ± 0.01
SMOTE(5, 30)	0.265 ± 0.02	0.524 ± 0.03	0.325 ± 0.03	0.535 ± 0.05	0.379 ± 0.02	0.526 ± 0.07	0.381 ± 0.01
SMOTE(7, 1)	0.672 ± 0.03	0.606 ± 0.03	0.710 ± 0.04	**0.997 ± 0.00**	0.495 ± 0.01	0.497 ± 0.00	0.497 ± 0.01
SMOTE(7, 3)	0.618 ± 0.03	0.596 ± 0.03	0.622 ± 0.04	**0.997 ± 0.00**	0.539 ± 0.03	0.497 ± 0.00	0.439 ± 0.04
SMOTE(7, 5)	0.557 ± 0.03	0.588 ± 0.03	0.594 ± 0.03	0.881 ± 0.08	0.526 ± 0.03	0.497 ± 0.00	0.403 ± 0.03
SMOTE(7, 10)	0.455 ± 0.03	0.568 ± 0.03	0.586 ± 0.03	0.628 ± 0.02	0.479 ± 0.03	0.497 ± 0.00	0.388 ± 0.01
SMOTE(7, 15)	0.377 ± 0.03	0.553 ± 0.03	0.578 ± 0.03	0.616 ± 0.02	0.443 ± 0.02	0.520 ± 0.06	0.385 ± 0.01
SMOTE(7, 20)	0.331 ± 0.02	0.539 ± 0.03	0.485 ± 0.03	0.604 ± 0.02	0.417 ± 0.02	0.515 ± 0.05	0.383 ± 0.01
SMOTE(7, 30)	0.268 ± 0.02	0.524 ± 0.03	0.333 ± 0.03	0.542 ± 0.04	0.379 ± 0.02	0.515 ± 0.05	0.381 ± 0.01
SMOTE(10, 1)	0.673 ± 0.03	0.606 ± 0.03	**0.711 ± 0.04**	**0.997 ± 0.00**	0.500 ± 0.02	0.497 ± 0.00	0.497 ± 0.01
SMOTE(10, 3)	0.621 ± 0.03	0.595 ± 0.03	0.626 ± 0.03	**0.997 ± 0.00**	**0.542 ± 0.03**	0.497 ± 0.00	0.445 ± 0.04
SMOTE(10, 5)	0.557 ± 0.03	0.588 ± 0.02	0.597 ± 0.03	0.865 ± 0.08	0.531 ± 0.03	0.497 ± 0.00	0.404 ± 0.03
SMOTE(10, 10)	0.458 ± 0.03	0.568 ± 0.02	0.587 ± 0.03	0.628 ± 0.02	0.479 ± 0.03	0.511 ± 0.05	0.389 ± 0.01
SMOTE(10, 15)	0.380 ± 0.03	0.554 ± 0.03	0.582 ± 0.04	0.615 ± 0.02	0.444 ± 0.02	0.512 ± 0.05	0.385 ± 0.01
SMOTE(10, 20)	0.331 ± 0.02	0.539 ± 0.03	0.489 ± 0.03	0.603 ± 0.02	0.420 ± 0.02	0.534 ± 0.07	0.384 ± 0.01
SMOTE(10, 30)	0.280 ± 0.03	0.524 ± 0.03	0.338 ± 0.03	0.542 ± 0.04	0.380 ± 0.01	0.526 ± 0.07	0.379 ± 0.01

Tablica 4.7. Rezultati G-mean mjere za klasifikacijski model GNB

Algoritam	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Baseline	0.682 ± 0.03	0.639 ± 0.03	0.500 ± 0.00	0.929 ± 0.04	0.500 ± 0.00	0.500 ± 0.00	0.500 ± 0.00
Poduzorkovanje	0.697 ± 0.03	**0.640 ± 0.03**	0.811 ± 0.03	**0.995 ± 0.01**	0.620 ± 0.06	0.495 ± 0.09	0.682 ± 0.07
SMOTE(3,1)	**0.701 ± 0.03**	0.637 ± 0.03	0.766 ± 0.05	**0.995 ± 0.01**	0.503 ± 0.01	0.500 ± 0.00	0.496 ± 0.02
SMOTE(3,3)	0.687 ± 0.03	0.635 ± 0.03	0.810 ± 0.03	**0.995 ± 0.01**	0.593 ± 0.06	0.500 ± 0.00	0.508 ± 0.05
SMOTE(3,5)	0.652 ± 0.02	0.630 ± 0.03	0.810 ± 0.02	0.969 ± 0.02	0.627 ± 0.06	0.500 ± 0.00	0.564 ± 0.08
SMOTE(3,10)	0.596 ± 0.02	0.620 ± 0.03	0.810 ± 0.02	0.884 ± 0.01	0.636 ± 0.07	0.508 ± 0.03	0.635 ± 0.09
SMOTE(3,15)	0.553 ± 0.02	0.609 ± 0.03	0.804 ± 0.02	0.875 ± 0.02	0.638 ± 0.06	0.528 ± 0.05	0.650 ± 0.09
SMOTE(3,20)	0.533 ± 0.01	0.602 ± 0.03	0.738 ± 0.03	0.869 ± 0.01	0.640 ± 0.05	0.534 ± 0.06	0.657 ± 0.10
SMOTE(3,30)	0.505 ± 0.01	0.594 ± 0.03	0.626 ± 0.02	0.812 ± 0.04	0.638 ± 0.05	**0.580 ± 0.09**	0.692 ± 0.09
SMOTE(5,1)	0.697 ± 0.03	0.638 ± 0.03	0.765 ± 0.04	**0.995 ± 0.01**	0.504 ± 0.02	0.500 ± 0.00	0.496 ± 0.02
SMOTE(5,3)	0.686 ± 0.02	0.635 ± 0.03	0.809 ± 0.03	**0.995 ± 0.01**	0.595 ± 0.05	0.500 ± 0.00	0.508 ± 0.05
SMOTE(5,5)	0.654 ± 0.02	0.630 ± 0.03	0.811 ± 0.02	0.971 ± 0.02	0.625 ± 0.05	0.500 ± 0.00	0.569 ± 0.08
SMOTE(5,10)	0.596 ± 0.02	0.621 ± 0.03	0.810 ± 0.02	0.885 ± 0.01	0.638 ± 0.07	0.500 ± 0.00	0.641 ± 0.09
SMOTE(5,15)	0.554 ± 0.02	0.610 ± 0.03	0.805 ± 0.02	0.876 ± 0.02	0.637 ± 0.06	0.515 ± 0.04	0.647 ± 0.09
SMOTE(5,20)	0.534 ± 0.01	0.601 ± 0.03	0.738 ± 0.03	0.870 ± 0.01	0.635 ± 0.06	0.528 ± 0.06	0.659 ± 0.10
SMOTE(5,30)	0.503 ± 0.01	0.594 ± 0.03	0.625 ± 0.02	0.820 ± 0.04	0.635 ± 0.05	0.538 ± 0.09	**0.694 ± 0.09**
SMOTE(7,1)	0.700 ± 0.03	0.638 ± 0.03	0.771 ± 0.05	**0.995 ± 0.01**	0.501 ± 0.01	0.500 ± 0.00	0.499 ± 0.00
SMOTE(7,3)	0.686 ± 0.03	0.635 ± 0.03	0.805 ± 0.03	**0.995 ± 0.01**	0.590 ± 0.05	0.500 ± 0.00	0.514 ± 0.06
SMOTE(7,5)	0.654 ± 0.02	0.631 ± 0.03	0.811 ± 0.02	0.973 ± 0.02	0.631 ± 0.06	0.500 ± 0.00	0.576 ± 0.07
SMOTE(7,10)	0.597 ± 0.02	0.620 ± 0.03	0.810 ± 0.02	0.885 ± 0.01	0.639 ± 0.07	0.500 ± 0.00	0.643 ± 0.09
SMOTE(7,15)	0.558 ± 0.02	0.609 ± 0.03	0.805 ± 0.02	0.877 ± 0.01	0.639 ± 0.06	0.515 ± 0.04	0.647 ± 0.09
SMOTE(7,20)	0.534 ± 0.01	0.602 ± 0.03	0.738 ± 0.02	0.869 ± 0.01	0.633 ± 0.06	0.515 ± 0.05	0.659 ± 0.10
SMOTE(7,30)	0.504 ± 0.01	0.594 ± 0.03	0.631 ± 0.02	0.825 ± 0.03	0.638 ± 0.05	0.527 ± 0.08	0.692 ± 0.09
SMOTE(10,1)	**0.701 ± 0.03**	0.638 ± 0.03	0.767 ± 0.04	**0.995 ± 0.01**	0.508 ± 0.02	0.500 ± 0.00	0.499 ± 0.00
SMOTE(10,3)	0.689 ± 0.02	0.634 ± 0.03	0.807 ± 0.03	**0.995 ± 0.01**	0.599 ± 0.06	0.500 ± 0.00	0.512 ± 0.05
SMOTE(10,5)	0.654 ± 0.02	0.631 ± 0.03	**0.812 ± 0.02**	0.969 ± 0.01	0.628 ± 0.06	0.500 ± 0.00	0.583 ± 0.09
SMOTE(10,10)	0.598 ± 0.02	0.620 ± 0.03	0.810 ± 0.02	0.884 ± 0.01	0.638 ± 0.07	0.508 ± 0.03	0.644 ± 0.09
SMOTE(10,15)	0.559 ± 0.02	0.610 ± 0.03	0.808 ± 0.02	0.876 ± 0.01	**0.641 ± 0.06**	0.510 ± 0.04	0.647 ± 0.09
SMOTE(10,20)	0.534 ± 0.01	0.601 ± 0.03	0.741 ± 0.03	0.868 ± 0.01	0.638 ± 0.06	0.526 ± 0.05	0.657 ± 0.10
SMOTE(10,30)	0.510 ± 0.02	0.594 ± 0.03	0.634 ± 0.02	0.825 ± 0.03	0.640 ± 0.05	0.537 ± 0.08	**0.694 ± 0.08**

Rezultati eksperimenta za klasifikacijski model SVM prikazani su u tablicama 4.8 i 4.9. Tablica 4.8 prikazuje prosječne vrijednosti F-1 mjere na 30 ponavljanja, dok tablica 4.9 prikazuje prosječne vrijednosti G-mean mjere uz pripadajuće standardne devijacije. Najviše prosječne vrijednosti za svaki skup podataka su prikazane podebljano.

Rezultati *baseline* pristupa su stabilni, no vrijednosti F-1 i G-mean su prilično niske za skupove s visokim omjerom neuravnoteženosti (D5-D7). Ipak, za skup D4 postiže se visoka prosječna vrijednost, slično kao kod drugih klasifikatora. Bitno je primjetiti da nijedan od skupova ne pokazuje najveću prosječnu vrijednost ni F-1 ni G-mean mjere za *baseline* pristup.

Rezultati pristupa nasumičnog poduzorkovanja slični su drugim klasifikatorima. F-1 vrijednosti su najčešće niže od *baseline* pristupa, osim u nekoliko slučajeva gdje su blago porasle. S druge strane, G-mean vrijednosti su veće od *baseline* pristupa, a za nekoliko skupova podataka su čak i najveće od svih pristupa.

Primjena SMOTE algoritma s različitim postavkama parametara daje nešto drukčije rezultate u odnosu na druge klasifikatore. Vrijednosti F-1 mjere su veće od *baseline* vrijednosti za sve skupove, a sličan obrazac se primjećuje i kod G-mean rezultata. Značajna je činjenica da su najveće vrijednosti i F-1 i G-mean bile postignute s kombinacijama parametara s većim vrijednostima parametra **k**, odnosno **k**=7 ili **k**=10, što razlikuje ovaj klasifikator od ostalih. Najbolja veličina parametra **q** opet je najviše ovisila o omjeru neuravnoteženosti pojedinih skupova podataka; skupovi s malim omjerom neuravnoteženosti imali su bolje performanse s manjim vrijednostima parametra **q**, dok su skupovi s visokim omjerima neuravnoteženosti imali bolje rezultate s većim vrijednostima parametra **q**. Bitno je napomenuti fenomen koji se pojavljuje kod skupa s najvećim omjerom neuravnoteženosti, D7. Naime, za vrijednosti parametra **q** ispod 15, rezultati ostaju gotovo nepromijenjeni. Ovo ukazuje na to da SMOTE, pri tako malim vrijednostima parametra **q**, nije dovoljno učinkovit u smanjenju problema neuravnoteženosti u slučaju SVM klasifikatora. Klasifikator se stoga ne uspijeva značajno poboljšati, što može biti posljedica nedovoljne promjene u distribuciji podataka koje SMOTE postiže pri manjim vrijednostima **q**.

SVM klasifikator, osobito kada se koristi u kombinaciji sa SMOTE-om, pokazuje značajan potencijal u radu s neuravnoteženim podacima. SMOTE može značajno poboljšati performanse modela, što je posebno vidljivo u rezultatima ovog eksperimenta.

Tablica 4.8. Rezultati F-1 mjere za klasifikacijski model SVM

Algoritam	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Baseline	0.713±0.03	0.659±0.03	0.866±0.02	0.989±0.01	0.492±0.00	0.524±0.07	0.498±0.00
Poduzorkovanje	**0.724±0.03**	0.719±0.02	0.833±0.02	**0.990±0.01**	0.470±0.02	0.495±0.06	0.368±0.05
SMOTE(3,1)	0.719±0.03	0.757±0.02	0.875±0.02	**0.990±0.01**	0.492±0.00	0.631±0.10	0.498±0.00
SMOTE(3,3)	0.587±0.03	0.726±0.03	0.777±0.03	0.988±0.01	0.513±0.03	0.667±0.11	0.498±0.00
SMOTE(3,5)	0.511±0.03	0.688±0.03	0.696±0.02	0.980±0.02	0.567±0.04	0.781±0.12	0.498±0.00
SMOTE(3,10)	0.408±0.03	0.619±0.03	0.565±0.02	0.871±0.04	0.544±0.03	0.886±0.09	0.498±0.00
SMOTE(3,15)	0.372±0.03	0.582±0.03	0.491±0.02	0.710±0.04	0.524±0.03	0.905±0.10	0.502±0.02
SMOTE(3,20)	0.355±0.03	0.560±0.03	0.436±0.02	0.621±0.03	0.501±0.03	0.916±0.09	0.508±0.02
SMOTE(3,30)	0.333±0.03	0.541±0.03	0.360±0.02	0.506±0.03	0.470±0.02	0.937±0.08	0.506±0.01
SMOTE(5,1)	0.720±0.03	0.757±0.03	0.874±0.02	0.989±0.01	0.492±0.00	0.651±0.11	0.498±0.00
SMOTE(5,3)	0.583±0.04	0.727±0.02	0.778±0.03	0.987±0.01	0.528±0.04	0.674±0.10	0.498±0.00
SMOTE(5,5)	0.507±0.04	0.684±0.03	0.698±0.03	0.978±0.03	0.567±0.04	0.772±0.14	0.498±0.00
SMOTE(5,10)	0.403±0.03	0.614±0.03	0.564±0.02	0.870±0.04	0.548±0.03	0.900±0.10	0.498±0.00
SMOTE(5,15)	0.364±0.03	0.576±0.03	0.488±0.02	0.711±0.04	0.522±0.03	0.916±0.09	0.503±0.02
SMOTE(5,20)	0.344±0.03	0.555±0.03	0.432±0.02	0.619±0.03	0.498±0.02	0.924±0.09	0.510±0.02
SMOTE(5,30)	0.325±0.02	0.534±0.04	0.358±0.02	0.502±0.03	0.464±0.02	0.945±0.07	0.509±0.01
SMOTE(7,1)	0.717±0.03	**0.758±0.03**	**0.876±0.02**	**0.990±0.01**	0.492±0.00	0.618±0.10	0.498±0.00
SMOTE(7,3)	0.582±0.04	0.726±0.02	0.777±0.03	0.988±0.01	0.519±0.04	0.687±0.10	0.498±0.00
SMOTE(7,5)	0.507±0.04	0.683±0.03	0.696±0.03	0.981±0.02	0.570±0.04	0.798±0.12	0.498±0.00
SMOTE(7,10)	0.398±0.03	0.611±0.03	0.563±0.02	0.866±0.04	0.550±0.03	0.904±0.09	0.498±0.00
SMOTE(7,15)	0.360±0.03	0.572±0.03	0.487±0.02	0.708±0.04	0.523±0.03	0.923±0.08	0.502±0.02
SMOTE(7,20)	0.340±0.02	0.552±0.03	0.432±0.02	0.620±0.03	0.501±0.02	0.934±0.08	**0.514±0.02**
SMOTE(7,30)	0.321±0.02	0.529±0.04	0.355±0.02	0.502±0.03	0.464±0.02	**0.947±0.07**	0.511±0.02
SMOTE(10,1)	0.717±0.03	0.755±0.03	0.876±0.02	**0.990±0.01**	0.492±0.00	0.615±0.11	0.498±0.00
SMOTE(10,3)	0.581±0.04	0.726±0.02	0.777±0.03	0.988±0.01	0.526±0.04	0.680±0.10	0.498±0.00
SMOTE(10,5)	0.506±0.04	0.683±0.03	0.697±0.02	0.981±0.02	**0.573±0.05**	0.831±0.10	0.498±0.00
SMOTE(10,10)	0.396±0.03	0.610±0.03	0.565±0.02	0.864±0.04	0.553±0.03	0.906±0.08	0.498±0.00
SMOTE(10,15)	0.356±0.03	0.572±0.03	0.488±0.02	0.707±0.04	0.523±0.03	0.930±0.08	0.501±0.02
SMOTE(10,20)	0.334±0.03	0.550±0.03	0.429±0.02	0.617±0.03	0.499±0.03	0.943±0.07	0.513±0.02
SMOTE(10,30)	0.320±0.02	0.525±0.04	0.353±0.02	0.502±0.03	0.462±0.02	**0.947±0.07**	0.513±0.02

Tablica 4.9. Rezultati G-mean mjere za klasifikacijski model SVM

Algoritam	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Baseline	0.704±0.03	0.640±0.03	0.847±0.03	0.982±0.01	0.500±0.00	0.517±0.04	0.500±0.00
Poduzorkovanje	0.738±0.03	0.782±0.03	**0.912±0.02**	**0.998±0.00**	**0.677±0.07**	0.777±0.14	**0.678±0.09**
SMOTE(3,1)	**0.748±0.03**	0.805±0.03	0.907±0.02	0.988±0.01	0.500±0.00	0.583±0.06	0.500±0.00
SMOTE(3,3)	0.669±0.03	0.818±0.02	0.900±0.02	0.991±0.01	0.511±0.02	0.613±0.08	0.500±0.00
SMOTE(3,5)	0.622±0.03	0.795±0.02	0.869±0.02	0.985±0.04	0.590±0.05	0.717±0.13	0.500±0.00
SMOTE(3,10)	0.564±0.02	0.747±0.02	0.786±0.02	0.960±0.06	0.653±0.07	0.833±0.12	0.500±0.00
SMOTE(3,15)	0.546±0.02	0.720±0.02	0.734±0.02	0.907±0.06	0.671±0.07	0.863±0.13	0.503±0.02
SMOTE(3,20)	0.538±0.02	0.705±0.02	0.693±0.02	0.858±0.07	0.662±0.07	0.875±0.12	0.530±0.05
SMOTE(3,30)	0.528±0.02	0.691±0.02	0.639±0.02	0.769±0.10	0.642±0.07	0.908±0.11	0.620±0.07
SMOTE(5,1)	**0.748±0.03**	0.804±0.03	0.906±0.03	0.988±0.01	0.500±0.00	0.600±0.08	0.500±0.00
SMOTE(5,3)	0.666±0.03	**0.819±0.02**	0.900±0.02	0.991±0.01	0.524±0.03	0.617±0.08	0.500±0.00
SMOTE(5,5)	0.620±0.03	0.792±0.02	0.870±0.02	0.983±0.04	0.588±0.06	0.713±0.13	0.500±0.00
SMOTE(5,10)	0.562±0.02	0.742±0.02	0.785±0.02	0.964±0.05	0.660±0.07	0.854±0.13	0.500±0.00
SMOTE(5,15)	0.543±0.02	0.716±0.02	0.731±0.02	0.910±0.06	0.670±0.07	0.875±0.12	0.505±0.02
SMOTE(5,20)	0.533±0.02	0.702±0.02	0.690±0.02	0.857±0.07	0.663±0.07	0.887±0.12	0.540±0.06
SMOTE(5,30)	0.523±0.01	0.687±0.03	0.639±0.02	0.767±0.10	0.632±0.08	0.917±0.10	0.644±0.10
SMOTE(7,1)	0.747±0.03	0.805±0.03	0.908±0.03	0.988±0.01	0.500±0.00	0.575±0.06	0.500±0.00
SMOTE(7,3)	0.665±0.03	0.818±0.02	0.901±0.02	0.990±0.01	0.517±0.03	0.625±0.07	0.500±0.00
SMOTE(7,5)	0.620±0.03	0.791±0.02	0.869±0.02	0.987±0.03	0.596±0.06	0.733±0.12	0.500±0.00
SMOTE(7,10)	0.559±0.02	0.741±0.02	0.786±0.02	0.959±0.06	0.664±0.06	0.858±0.12	0.500±0.00s
SMOTE(7,15)	0.540±0.02	0.714±0.02	0.731±0.02	0.906±0.06	0.674±0.07	0.883±0.12	0.505±0.02
SMOTE(7,20)	0.531±0.01	0.700±0.02	0.691±0.02	0.859±0.07	0.666±0.07	0.900±0.11	0.543±0.04
SMOTE(7,30)	0.522±0.01	0.684±0.02	0.637±0.02	0.767±0.10	0.634±0.08	**0.921±0.10**	0.645±0.09
SMOTE(10,1)	0.747±0.03	0.802±0.03	0.907±0.02	0.988±0.01	0.500±0.00	0.575±0.07	0.500±0.00
SMOTE(10,3)	0.665±0.03	**0.819±0.02**	0.901±0.02	0.992±0.01	0.522±0.03	0.621±0.08	0.500±0.00
SMOTE(10,5)	0.619±0.03	0.792±0.02	0.869±0.02	0.986±0.03	0.598±0.06	0.767±0.12	0.500±0.00
SMOTE(10,10)	0.558±0.02	0.740±0.02	0.787±0.02	0.956±0.06	0.660±0.07	0.858±0.11	0.500±0.00
SMOTE(10,15)	0.538±0.02	0.714±0.02	0.732±0.02	0.907±0.06	0.671±0.07	0.892±0.11	0.502±0.02
SMOTE(10,20)	0.528±0.02	0.698±0.02	0.688±0.02	0.856±0.07	0.658±0.07	0.912±0.10	0.541±0.05
SMOTE(10,30)	0.522±0.01	0.681±0.03	0.635±0.02	0.766±0.10	0.631±0.08	**0.921±0.10**	0.644±0.08

Analizirani klasifikatori pokazali su različite razine osjetljivosti na neuravnoteženost podataka i primjenu različitih metoda uravnoteženja klasa. GNB i SVM modeli su se posebno istaknuli u primjeni SMOTE algoritma, koji je u velikoj mjeri poboljšao njihove performanse. Međutim, za sve klasifikatore bilo je ključno pravilno podešavanje parametara kako bi se postigla najučinkovitija ravnoteža između osjetljivosti i specifičnosti. Unatoč određenim ograničenjima, SMOTE se pokazao kao najbolja tehnika u borbi protiv neuravnoteženosti podataka, dok su poduzorkovanje i *baseline* pristupi imali ograničene učinke, posebno u situacijama s visokom razinom neuravnoteženosti klasa. Rezultati ovih eksperimenata naglašavaju važnost prilagodbe parametara SMOTE algoritma specifičnostima pojedinih skupova podataka kako bi se postigle najbolje moguće performanse.

Za SMOTE algoritam najefikasnije vrijednosti parametara ovisile su o više različitih faktora:

- **Omjer neuravnoteženosti skupa podataka** – najučinkovitija vrijednost parametra **q** najviše je ovisila o omjeru neuravnoteženosti (IR) skupa podataka; za visoke IR obično je vrijedilo da su učinkovitije visoke vrijednosti parametra **q**, dok za niske vrijednosti IR je vrijedilo da su učinkovitije niske vrijednosti **q**.
- **Klasifikator** – provođenje eksperimentalne analize na nekoliko različitih klasifikatora pokazalo je da iste postavke parametara algoritma SMOTE nemaju isti učinak na svakom klasifikatoru, što jasno upućuje na zaključak da najučinkovitije postavke parametara uvelike ovise i o klasifikatoru koji se koristi pri klasifikaciji podataka.
- **Veličina skupa podataka** – najbolje vrijednosti parametara SMOTE-a mogu također ovisiti o veličini skupa podataka. Tako je najučinkovitija vrijednost parametra **k** za brojčanije skupove češće bila veća nego za one manje brojne.

5. ZAKLJUČAK

Zadatak ovog završnog rada bio je istražiti utjecaj različitih postavki parametara algoritma SMOTE na performanse klasifikacijskih modela primjenjenih na neuravnotežene skupove podataka. Tijekom rada, provedena je detaljna analiza problema neuravnoteženosti klasa, klasifikacijskih modela te evaluacijskih mjera, s posebnim fokusom na algoritam SMOTE. Kroz eksperimentalnu analizu ispitani su učinci različitih postavki ključnih parametara SMOTE algoritma, uključujući broj susjeda (**k**) i broj sintetičkih uzoraka (**q**), na performanse klasifikatora poput 5-NN, stabla odluke, GNB i SVM.

Eksperimentalna analiza pokazala je da pažljivo odabrane postavke parametara SMOTE algoritma mogu značajno poboljšati performanse klasifikacijskih modela, posebno kada postoji visok omjer neuravnoteženosti među klasama. Međutim, rezultati također upućuju na to da treba oprezno postavljati parametre kako bi se izbjeglo pretjerano generiranje sintetičkih podataka, što može dovesti do prekomjerne prilagodbe modela podacima za treniranje. Prekomjerna prilagodba može dovesti do smanjenja F-1 i G-Mean mjera, jer model postaje previše fokusiran na specifične obrasce koji ne odražavaju stvarnu distribuciju podataka. Stoga je važno pronaći optimalnu ravnotežu između generiranja dovoljne količine sintetičkih podataka za poboljšanje klasifikacije manjinskih klasa i očuvanja sposobnosti modela za generalizaciju.

Eksperimenti su također pokazali da postavke algoritma SMOTE treba prilagoditi specifičnim karakteristikama skupa podataka i klasifikacijskog modela, jer ne postoji univerzalno rješenje koje uvijek daje najbolje rezultate. U određenim slučajevima, primjena SMOTE-a može biti suvišna, kao kod skupova podataka na kojim klasifikatori već postižu dobre performanse bez preuzorkovanja. Također, za neke klasifikatore, poput stabla odluke, utjecaj SMOTE algoritma bio je minimalan ili čak negativan na performanse modela na većini ispitanih skupova podataka.

Kao prijedlog za budući rad, bilo bi korisno istražiti utjecaj parametara na druge varijante SMOTE algoritma, poput *Borderline-SMOTE* ili *ADASYN*, te dodatno analizirati kako različiti klasifikatori reagiraju na te varijante i njihove različite postavke parametara. Također, proširenje eksperimentalne analize na višeklasne probleme moglo bi pružiti dublji uvid u prilagodljivost SMOTE algoritma u složenijim okruženjima.

LITERATURA

[1] N. Japkowicz, M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[2] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2000.

[3] S. Theodoridis, K. Koutroumbas, *Pattern Recognition, Fourth Edition*, izd. 4, Academic Press, 2008.

[4] K. Odajima, A. P. Pawlovsky, „A detailed description of the use of the knn method for breast cancer diagnosis“, *Proceedings of the 7th International Conference on BioMedical Engineering and Informatics (BMEI'14)*, str. 668–692, 2014.

[5] H. Kamel, D. Abdulah, J. M. Al-Tuwaijari, „Cancer classification using gaussian naive bayes algorithm“, *Proceedings of the 5th International Engineering Conference on Developments in Civil & Computer Engineering Applications (IEC'19)*, str. 165–170, 2019.

[6] V. Lopez, A. Fernandez, S. Garcia, V. Palade, F. Herrera, „An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics“, *Information Sciences*, sv. 250, str. 113–141, 2013.

[7] D. Bajer, B. Zorić, M. Dudjak, G. Martinović, „Performance analysis of smote-based oversampling techniques when dealing with data imbalance“, *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, str. 265–271, 2019.

[8] M. Dudjak, G. Martinović, „In-depth performance analysis of smote-based over-sampling algorithms in binary classification“, *International journal of electrical and computer engineering systems*, izd. 1, sv. 11, str. 13–23, 2020.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, „SMOTE: Synthetic minority over-sampling technique“, *Journal Of Artificial Intelligence Research*, sv. 16, str. 321–357, 2002.

[10] J. Stefanowski, S. Wilk, „Improving rule based classifiers induced by modlem by selective pre-processing of imbalanced data“, *Proceedings of the 18th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'07)*, str. 54–65, 2007.

[11] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, „Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem“, *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD'09)*, str. 475–482, 2009.

[12] C. Bellinger, C. Drummond, N. Japkowicz, „Beyond the boundaries of SMOTE“, *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'16)*, str. 248–263, 2016.

[13] B. Zorić, D. Bajer, G. Martinović, „Employing different optimisation approaches for SMOTE parameter tuning“, *Proceedings of the 2016 International Conference on Smart Systems and Technologies (SST'16)*, str. 191–196, 2016.

[14] A. Fernandez, S. Garcia, F. Herrera, N. V. Chawla, „SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary“, *Journal Of Artificial Intelligence Research*, sv. 61, str. 863–905, 2018.

[15] H. Han, W.-Y. Wang, B.-H. Mao, „Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning“, *Proceedings of the International Conference on Intelligent Computing* (*ICIC'05)*, str. 878–887, 2005.

[16] Y. Dong, X. Wang, „A new over-sampling approach: Random SMOTE for learning from imbalanced data sets“, *Proceedings of the 5th International Conference on Knowledge Science, Engineering and Management (KSEM'11)*, str. 343–352, 2011.

SAŽETAK

U radu je istražen problem skupova podataka sa svojstvom neuravnoteženosti klasa koje otežava klasifikaciju manjinske klase, te mogućnosti rješavanja tog problema primjenom algoritma SMOTE, koji pruža preuzorkovanje podataka stvaranjem sintetičkih uzoraka manjinske klase s ciljem postizanja ravnoteže između klasa. Posebna pažnja posvećena je analizi utjecaja različitih postavki parametara algoritma SMOTE, uključujući broj susjeda i broj generiranih sintetičkih uzoraka, na performanse klasifikacijskih modela. Eksperimentalna analiza provedena je koristeći više klasifikatora, pri čemu su rezultati ocjenjivani F-1 i G-Mean mjerama. Pokazalo se da pažljivo odabrane postavke parametara mogu značajno poboljšati učinkovitost klasifikacije, dok nepravilne postavke mogu dovesti do prekomjerne prilagodbe podacima za treniranje i smanjenja performansi. Zaključeno je da se parametri algoritma SMOTE trebaju prilagoditi specifičnostima skupa podataka i klasifikatora kako bi se postiglo poboljšanje učinkovitosti klasifikacije.

Ključne riječi: klasifikacija, neuravnoteženost klasa, parametri algoritma SMOTE, preuzorkovanje, sintetički uzorci

ABSTRACT

Impact of the parameters of the SMOTE algorithm for handling class imbalance

The paper explores the issue of datasets with class imbalance, which complicates the classification of the minority class, and the possibilities of addressing this problem using the SMOTE algorithm , which enables data resampling by creating synthetic samples of the minority class with the aim of achieving balance between classes.. Special attention is given to analyzing the impact of different settings of SMOTE algorithm parameters, including the number of neighbors and the number of generated synthetic samples, on the performance of classification models. The experimental analysis was conducted using multiple classifiers, with results evaluated by the F-1 and G-Mean metrics. It was shown that carefully selected parameter settings can significantly improve classification performance, while incorrect settings can lead to overfitting to the training data and reduce performance. It is concluded that the parameters of the SMOTE algorithm should be adjusted to the specific characteristics of the dataset and the classification model to achieve classification efficiency improvements.

Keywords: classification, class imbalance, SMOTE algorithm parameters, oversampling, synthetic samples