Optimizacija procesa odabira značajki u strojnom učenju

Kumir, Marko

SVEUČILIŠTE U SPLITU
PRIRODOSLOVNO MATEMATIČKI FAKULTET

DIPLOMSKI RAD

OPTIMIZACIJA PROCESA ODABIRA ZNAČAJKI U STROJNOM UČENJU

Marko Kumir

Split, rujan 2024.

Temeljna dokumentacijska kartica

Diplomski rad

Sveučilište u Splitu
Prirodoslovno-matematički fakultet
Odjel za informatiku
Ruđera Boškovića 33, 21000 Split, Hrvatska

Optimizacija procesa odabira značajki u strojnom učenju

Marko Kumir

SAŽETAK

Pravilno odabrane značajke igraju ključnu ulogu u poboljšanju performansi modela strojnog učenja, smanjenju složenosti podataka i ubrzavanju procesa treniranja. Nakon detaljnog pregleda literature provedbom PRISMA smjernica, fokus je stavljen na financijsku domenu, posebno na procjenu kreditnog rizika u bankarstvu. U eksperimentalnom dijelu rada testirane su različite filter metode i ugrađene metode, pri čemu je testiranje provedeno na tri kreditna skupa podataka. S ciljem optimizacije procesa, testiranje je izvršeno na 50% najutjecajnijih značajki. Rezultati su pokazali da su RF i XGB modeli, trenirani na reduciranim skupovima podataka, ostvarili visoke performanse u usporedbi s modelima treniranim na inicijalnim skupovima podataka.

Ključne riječi: odabir značajki, strojno učenje, kreditni rizik, filter metode, ugrađene metode, SMOTE, target encoding, redukcija dimenzionalnosti

Rad je pohranjen u knjižnici Prirodoslovno-matematičkog fakulteta, Sveučilišta u Splitu

Rad sadrži: 41 stranicu, 5 grafičkih prikaza, 7 tablica i 22 literaturna navoda. Izvornik je na hrvatskom jeziku.

Mentor: **Dr. sc. Goran Zaharija,** *docent, Prirodoslovno-matematički fakultet, Sveučilište u Splitu*

Ocjenjivači: **Dr. sc. Goran Zaharija,** *docent, Prirodoslovno-matematički fakultet, Sveučilište u Splitu*

Dr. sc. Divna Krpan, *docent, Prirodoslovno-matematički fakultet, Sveučilište u Splitu*

Nika Jerković, *asistent, Prirodoslovno-matematički fakultet, Sveučilište u Splitu*

Rad je prihvaćen: **rujan, 2024.**

Basic documentation card

Graduate thesis

University of Split
Faculty of Science
Department of Computer Science
Ruđera Boškovića 33, 21000 Split, Croatia

Optimization of Feature Selection Process in Machine Learning

Marko Kumir

ABSTRACT

Properly selected features play a crucial role in improving the performance of machine learning models, reducing data complexity, and speeding up the training process. After a detailed literature review conducted using PRISMA guidelines, the focus was placed on the financial domain, specifically on credit risk assessment in banking. In the experimental part of the study, various filter methods and embedded methods were tested, with experiments conducted on three credit datasets. To optimize the process, testing was carried out on 50% of the most relevant features. The results showed that the RF and XGB models, trained on reduced datasets, achieved high performance compared to models trained on the initial datasets.

Key words: feature selection, machine learning, credit risk, filter methods, embedded methods, SMOTE, target encoding, dimensionality reduction

Thesis deposited in library of Faculty of science, University of Split

Thesis consists of: 41 pages, 5 figures, 7 tables and 22 references

Original language: Croatian

Mentor: **Goran Zaharija, Ph.D.** *Assistant Professor of Faculty of Science, University of Split*

Reviewers: **Goran Zaharija, Ph.D.** *Assistant Professor of Faculty of Science, University of Split*

Divna Krpan, Ph.D. *Assistant Professor of Faculty of Science, University of Split*

Nika Jerković *Instructor of Faculty of Science, University of Split*

Thesis accepted: **September, 2024.**

IZJAVA

kojom izjavljujem s punom materijalnom i moralnom odgovornošću da sam diplomski rad s naslovom „*Optimizacija procesa odabira značajki u strojnom učenju*“ izradio samostalno pod voditeljstvom dr.sc. Gorana Zaharije. U radu sam primijenio metodologiju znanstvenoistraživačkog rada i koristio literaturu koja je navedena na kraju diplomskog rada. Tuđe spoznaje, stavove, zaključke, teorije i zakonitosti koje sam izravno ili parafrazirajući naveo u diplomskom radu na uobičajen, standardan način citirao sam i povezao s fusnotama s korištenim bibliografskim jedinicama. Rad je pisan u duhu hrvatskog jezika.

Student:

Marko Kumir

Sadržaj

Uvod

U kontekstu strojnog učenja, kvaliteta i broj značajki koje se koriste za treniranje modela igraju ključnu ulogu u postizanju točnosti i učinkovitosti modela. Proces odabira značajki (engl. *Feature Selection*) predstavlja korak kojim se optimizira, odnosno smanjuje broj značajki u skupu podataka, zadržavajući pritom samo one značajke koje najviše doprinose prediktivnim sposobnostima modela. Pravilno odabran skup značajki može značajno poboljšati performanse modela, smanjiti mogućnost prekomjernog prilagođavanja (engl. *overfitting*), pojednostavniti interpretaciju rezultata te ubrzati proces treniranja.

Prije provedbe eksperimentalnog dijela, realizirane su PRISMA (engl. *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) smjernice kako bi se omogućio detaljan pregled istraženosti područja odabira značajki u strojnom učenju. Početno istraživanje obuhvatilo je širok spektar primjena odabira značajki u različitim domenama. Finalan fokus analize usmjeren je na financijsku domenu, specifično na radove koji se bave procjenom kreditnog rizika u bankarstvu, gdje je odabir relevantnih značajki od iznimne važnosti zbog složenosti i visoke dimenzionalnosti ulaznih podataka. Financijske institucije često se suočavaju s izazovom donošenja odluka o odobravanju ili odbijanju kreditnih zahtjeva, pri čemu postoji rizik da se klijent koji dobije odobrenje ne pridržava ugovorenih obveza plaćanja, ili da se odbije zahtjev klijenta koji može izvršavati svoje obveze plaćanja bez kašnjenja. Pogrešno procjene često rezultiraju financijskim gubitcima.

U eksperimentalnom dijelu rada korištena su tri kreditna skupa podataka s ciljem izgradnje modela za procjenu kreditnog rizika. U zadnjoj fazi pretprocesiranja podataka, implementirane su četiri metode za odabir značajki koje uključuju univarijantnu filter metodu Mutual Information, multivarijantnu filter metodu Pearson Correlation, te ugrađene (engl. *embedded*) metode relevantne za modele Random Forest i XGB. Navedene metode su primijenjene nad svim skupovima podataka, kako bi se testirala njihova učinkovitost i usporedila s performansama modela nad inicijalnim skupovima podataka. Korišteni su modeli strojnog učenja Random Forest i XGB koji su testirani pomoću evaluacijskih metrika Accuracy, F1-score i AUC, čime se osigurala sveobuhvatna procjena njihovih izvedbi.

1. Pregled istraženosti područja

U ovom radu proveden je sustavni pregled literature (engl. *Systematic Literature Review*) s ciljem analize suvremenih metoda za odabir značajki u strojnom učenju. Sustavni pregled literature je istraživačka metodologija za prikupljanje, identifikaciju i kritičku analizu dostupnih istraživačkih studija putem sustavne procedure[1]. Pruža sveobuhvatan pregled postojećih rješenja, tehnika i primjena u određenom znanstvenom području.

1.1. Strategija pretraživanja

Prvi korak ovog procesa je bio pretraživanje literature kako bi se identificirale sve objavljene studije u kontekstu strojnog učenja koje se bave implementacijom procesa odabira značajki u posljednje dvije godine. Za pretragu su korištene dvije baze podataka: Web of Science i Scopus. Ove baze podataka su odabrane iz razloga što nude širok spektar filtera i kvalitetne publikacije. Korišteni su sljedeći filteri za pretragu:

- Web of Science → TI=(Feature selection OR Variable selection) AND AK=(machine learning) AND SU=(Computer Science) AND DT=(Article) AND PY=(2023-2024) AND LA=(English)
- Scopus → TITLE (Feature selection OR Variable selection) AND KEY (machine learning) AND SUBJAREA (comp) AND DOCTYPE (ar) AND PUBYEAR > 2022 AND LANGUAGE (english)

Specifični filteri pretrage
Radovi unutar znanstvenog područja računarstva
Recenzirani radovi klasificirani kao "Article"
Radovi objavljeni nakon 2022. godine
Radovi objavljeni na engleskom jeziku

Tablica 1 Opis specifičnih filtera pretrage

Članci su pretraživani po naslovu, ključnim riječima, znanstvenom području, tipu dokumenta, godini objavljivanja i jeziku pisanja. Razlog odabira članaka objavljenih nakon 2022. godine leži u činjenici da ovo područje obuhvaća širok spektar istraživanja, te je naglasak stavljen isključivo na najnovije radove kako bi se osigurala relevantnost i ažurnost informacija.

Rezultati primjenom filtera su pokazali ukupno 798 potencijalno relevantnih članaka, odnosno WOS je rezultirao s 240, a Scopus s 558 radova. Nakon provjere duplikata, odbačen je 201 rad, odnosno identificirano je 597 jedinstvenih radova koji su prešli iz faze identifikacije u fazu pregledavanja (engl. *screening*).

1.2. Kriterije isključivanja i uključivanja

Tijekom faze pregledavanja jedinstvenih, identificiranih radova, primijenjeni su specifični kriteriji isključivanja i uključivanja u svrhu pronalaska prikladnih finalnih članaka za daljnju analizu.

Kriteriji isključivanja:

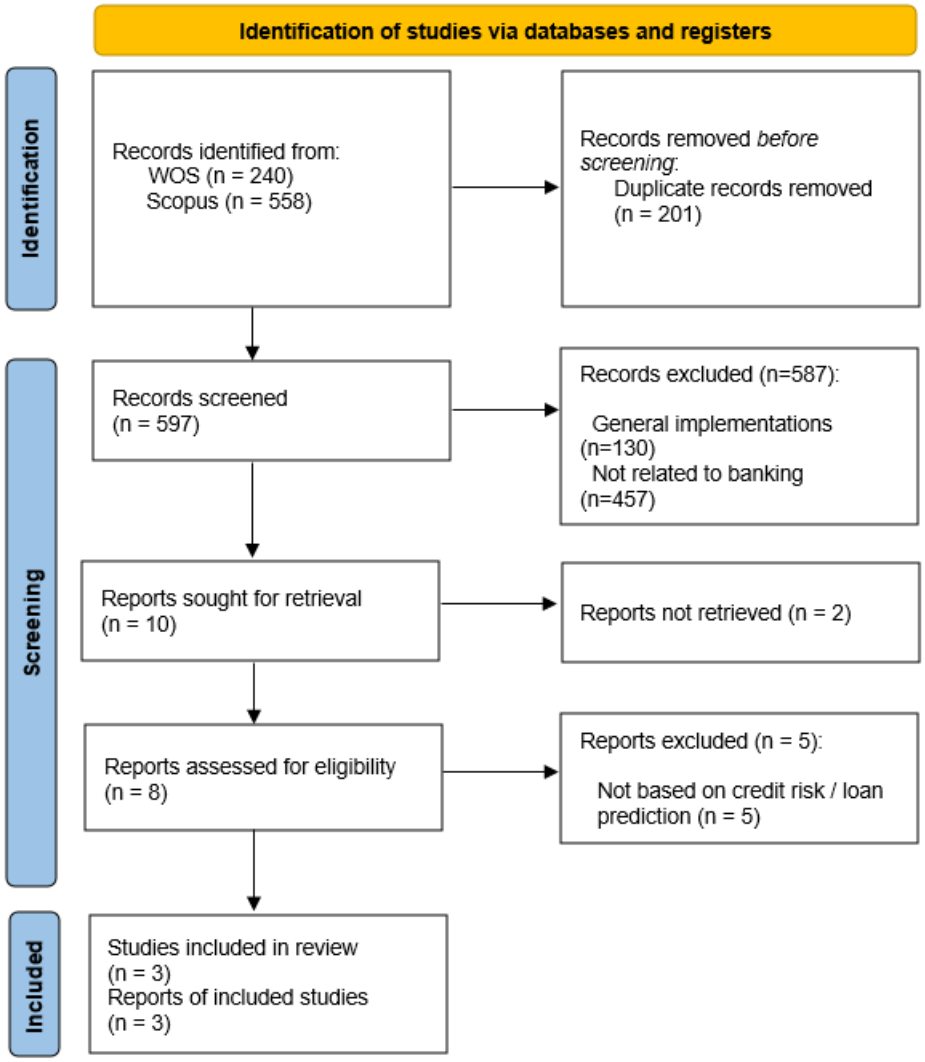
- Generalne implementacije algoritama odabira značajki
- Primjena algoritama odabira značajki na specifične probleme van domene bankarstva

130 radova je odbačeno jer su se bavili općim implementacijama algoritama za odabir značajki, stavljajući fokus na više domena. Drugi kriterij isključio je dodatnih 457 radova koji su kroz primjenu algoritama za odabir varijabli evaluirali modele za specifičnu predikciju koja nije bila vezana uz bankarsku domenu. Ukupno, nakon realizacije kriterija isključivanja, 587 radova je isključeno iz daljnje analize.

Specifični kriterij uključivanja:

- Radovi bazirani na predikciji u kontekstu kreditnog rizika

U fazi dohvaćanja (engl. *retrieval*) fazi provjerena je dostupnost cjelovitih radova. 2 rada su isključena jer je bio dostupan samo sažetak, što je rezultiralo s 8 radova u fazi podobnosti (engl. *eligibility*) koji su detaljno analizirani. Primjenom specifičnog kriterija uključivanja 5 radova je odbačeno zbog nezadovoljavajućeg predmeta predikcije. Na kraju, 3 rada su ispunila specifičan kriterij i korišteni su za daljnju analizu u ovom pregledu literature.



Slika 1 PRISMA dijagram toka

1.3. Meta-analiza

Ovo poglavlje obuhvaća detaljnu analizu odabranih radova. Cilj je usporediti radove kroz glavne aspekte, uključujući korištene skupove podataka, fazu pretprocesiranja podataka, odabir modela te evaluacije odabranih modela nad odabranim skupovima značajki. Važno je napomenuti da je fokus ove analize na pregledu načina implementacija metoda za odabir značajki prije faze treniranja modela.

Uključeni radovi imaju slične ciljeve unutar konteksta procjene kreditnog rizika. [2] Prvi članak se fokusira na predikciju hoće li klijent prihvatiti ili odbiti ponudu osobnog kredita kojeg mu je banka unaprijed odobrila, [3] drugi članak se bavi procjenom klasifikacije

„dobrih“ i „loših“ klijenata koji su već dobili kredit, dok [4] treća odabrana publikacija predviđa dva slučaja, a to su odobravanje/odbijanje kredita na temelju kreditne sposobnosti i klasifikacija klijenata s dobivenim kreditima na klijente koji su ispunili obveze (engl. *non-default clients*) u dogovorenom roku i na one koji nisu (engl. *default clients*).

1.3.1. Skupovi podataka

Glavne specifikacije korištenih skupova podataka odabranih studija su navedene u tablici ispod.

Rad	Dataset (izvor)	Broj podataka	Broj značajki
[2]	Thera Bank dataset (Kaggle)	5000	14
[3]	Lending Club dataset (Kaggle)	1.408.575	152
[4]	Australian dataset (UCI)	690	14
[4]	German dataset (UCI)	1000	20
[4]	Taiwan dataset (UCI)	30.000	24

Tablica 2 Specifikacije skupova podataka odabranih publikacija

Iz priložene tablice vidimo da je rad [4] za predikciju u području kreditnog rizika koristio tri skupa podataka, dok su preostala dva rada koristili po jedan skup podataka. Rad [3] prednjači s inicijalnim dimenzijama korištenog skupa podataka za razliku od ostalih radova.

Kada se uzmu u obzir detaljnije specifikacije, uključujući nedostajuće vrijednosti i balansiranost podataka, navedeno je sljedeće: Lending Club skup podataka sadrži znatnu količinu nedostajućih vrijednosti, dok Thera Bank nema nedostajućih podataka. Za preostala tri skupa podataka ova informacija nije specifično navedena. Što se tiče balansiranosti podataka, odnosno omjera pozitivne i negativne klase, skupovi podataka Thera Bank, Lending Club i Taiwan su poprilično neuravnoteženi, dok su Australian i German skupovi podataka relativno uravnoteženi.

1.3.2. Pretprocesiranje podataka

S obzirom da skup podataka rada [2] ne sadrži nedostajuće vrijednosti, jedini korak prije provedbe metoda za odabir varijabli bio je manualno uklanjanje ID i Zip Code atributa iz skupa značajki, jer nisu bili relevantni za kreditnu procjenu. Nakon inicijalnog reduciranja dimenzionalnosti, proveden je odabir značajki pomoću šest različitih metoda: Pearson, Chi-2, RFE, Logistic Regression (pomoću regularizacije), Random Forest i LightGBM (na temelju važnosti značajki). Kombinacijom metoda izabran je hibridni skup značajki koji se sastojao od 6 varijabli: Education, Income, Family, CD Account, CCAvg i Experience. Iako u radu [2] nije bilo izričito obrazloženo zašto je odabrano točno šest navedenih značajki, analizom rezultata odabira značajki hibridnim pristupom može se zaključiti da je svaka varijabla odabrana pomoću najmanje četiri različite metode. Ostale značajke nisu zadovoljavale taj kriterij, jer su bile odabrane preko tri ili manje metoda. Može se pretpostaviti da je inicijalno postavljen uvjet zadovoljavajućeg broja metoda koji je osigurao odabir konzistentnijih značajki.

Faza pretprocesiranja podataka drugog rada [3] je bila dosta kompleksnija za razliku od prethodnog rada s naglaskom na to da je skup podataka ovog rada bio i znatnije većih dimenzija. Prvi korak ove faze je bio manualno uklanjanje irelevantnih značajki. Prema mišljenju stručnjaka uklonjena su 53 atributa koji su procijenjeni kao nebitni za predikciju kreditne sposobnosti klijenta. Drugi korak je bio rukovanje s nedostajućim vrijednostima unutar kojeg su sve značajke s više od 50% nedostajućih vrijednosti eliminirane iz daljnjeg razmatranja. Preostalim varijablama su nedostajuće vrijednosti zamijenjene prosječnom vrijednosti stupca u slučaju numeričkog tipa podataka, dok su kod značajki sa kategoričkim tipom podataka zamijenjene s modom stupca. Realiziranjem ovog postupka preostale su 33 značajke. Potom su kategoričke varijable enkodirane u numeričke vrijednosti, te su podatci standardizirani. Poslije toga slijedila je provjera korelacije među nezavisnim značajkama kojom je analizirana korelacija između parova nezavisnih varijabli kako bi se identificirale visoko povezane varijable, preciznije parovi sa korelacijom iznad praga od 0,8. Iz parova koji su prelazili postavljeni prag, jedna značajka je uklonjena kako bi se smanjila redundancija i samim tim izbjegao problem multikolinearnosti. Implementacijom ovog procesa uklonjeno je 8 redundantnih značajki. Zadnji korak je bila provjera korelacija s ciljnom varijablom. Primijenjen je prag korelacije od 0,97 na nezavisne varijable sa predviđanom kako bi se izbacile izrazito visoko korelirane nezavisne značajke.

Implementacijom pretposljednje faze pretprocesiranja odbačeno je šest varijabli čime preostaje 19 značajki koje su korištene u zadnjoj fazi.

U fazi primjene metoda za odabir značajki korištene su četiri tehnike: UFS, RFE, IV, FIDT. Svaka od metoda je odabirala 9 najrelevantnijih varijabli za treniranje modela. Detaljnom analizom utvrđeno je da su nakon primjene svake tehnike, na temelju odabranih varijabli, kreirani odgovarajući skupovi podataka za treniranje i testiranje koji su naknadno upotrijebljeni za treniranje i evaluaciju odabranog modela.

U zadnjem radu [4] faza pretprocesiranja podataka se sastojala od faze normalizacije podataka i faze primjene metoda za odabir značajki. Nije korišten hibridni pristup više metoda kao što je to bila primjena u prethodna dva rada već je korištena jedna tehnika, a to je Information Gain. Autori ovog rada su naveli da je razlog odabira ove filter tehnike to što je računalno manje zahtjevna u usporedbi s wrapper i hibridnim pristupima. Ova metoda je provedena kroz sva tri korištena skupa podataka. Australski skup podataka je reduciran s 14 na 9 varijabli, njemački s 20 na 13, a tajvanski s 24 na 17 značajki. Nakon identifikacije najbitnijih značajki, napravljena je podjela reduciranih skupova podataka na podatke za treniranje i testiranje koji su služili za daljnju fazu modeliranja. Svi skupovi podataka su testirani zasebno i njihovi rezultati su naknadno uspoređeni.

Rad	FS metode	Broj značajki prije FS	Broj značajki nakon FS	Postotak redukcije
[2]	Hybrid (Pearson, Chi-2, RFE, LR, RF, LightGBM)	14	6	57,14 %
[3]	Chi-2, RFE, FIDT, IV	19	9	52,63 %
[4]	Information Gain	14 (Australian) 20 (German) 24 (Taiwannese)	9 (Australian) 13 (German) 17 (Taiwannese)	35,71 % 35,00 % 29,17 %

Tablica 3 Usporedba metoda za odabir značajki u odabranim publikacijama

1.3.3. Modeliranje i evaluacija

Uspoređene su i faze modeliranja i evaluacije odabranih modela u svakom od radova kako bi se istaknule sličnosti i razlike u metodologijama, korištenim modelima i evaluacijskim metrikama.

U prvom radu [2] korišten je najširi spektar modela u odnosu na ostale radove. Korišteni modeli strojnog učenja su: Support Vector Machine (skraćeno SVM), Gaussian Naive Bayes (skraćeno GNB), Random Forest (skraćeno RF), Logistic Regression, AdaBoost, Decision Tree (skraćeno DT), k-Nearest Neighbors (skraćeno k-NN), Gradient Boosting Classifier (skraćeno GBC), Voting Classifier, i Stacking. Svi modeli su evaluirani nad originalnim podacima i na reduciranim podacima. Cilj je bio usporediti rezultate evaluacije modela nad podacima prije i poslije primjene metoda za odabir značajki, te procijeniti koliko primjena odabira značajki može poboljšati performanse različitih modela strojnog učenja.

U drugoj publikaciji [3] korišten je samo jedan model strojnog učenja, a to je Extreme Gradient Boosting (skraćeno XGB), koji je prikladan za treniranje i evaluaciju nad velikim skupovima podataka. XGB je treniran na četiri različita skupa podataka, pri čemu je svaki skup zapravo bio reduciran skup podataka koji je proizašao primjenom jedne od četiri metode odabira značajki. Nakon treniranja, svaki od četiri modela evaluiran je koristeći iste evaluacijske metrike: F1, AUC, Accuracy, Precision i Recall. Navedene metrike su korištene u svrhu usporedbe performansi XGB modela temeljenih na različitim skupovima značajki.

Zadnji članak [4] obrađuje sljedeća četiri modela strojnog učenja: Random Forest, Gradient Boosting, Extreme Gradient Boosting i Stacked Classifier. Svi modeli su evaluirani koristeći reduciran skup značajki dobivenih filter metodom Information Gain. Korišteni Stacked Classifier je sekvencijalni model koji se sastoji od više estimatora. Za razliku od pojedinačnih modela, „stacked“ pristup kombinira izlaze različitih modela kako bi generirao konačnu odluku. Kreirana su tri različito konfigurirana Stacked Classifier-a, od kojih je svaki treniran na određenom skupu podataka (australski, njemački, tajvanski). Odabrani modeli su evaluirani pomoću sljedećih evaluacijskih metrika: Accuracy, F1 i AUC. Primarni fokus evaluacije ovog rada bio je na usporedbi izvedbi svih korištenih modela na tri skupa podataka.

1.3.4. Rezultati

Autori rada [2] koji su koristili Thera Bank skup podataka su usporedili rezultate evaluacije modela prije i nakon primjene metoda odabira značajki. Iako je skup značajki bio značajno reduciran, rezultati evaluacija modela treniranim nad reduciranim skupovima ostali su gotovo nepromijenjeni u odnosnu na rezultate evaluacija modela nad početnim podacima. Random Forest i Stacking modeli su postigli najbolje rezultate u oba slučaja. Kod Lending Club skupa podataka u radu [3], XGB model je evaluiran na četiri različita skupa podataka koji su dobiveni primjenom različitih metoda odabira značajki navedenih u Tablici 3. XGB modeli trenirani nad reduciranim skupovima podataka, dobiveni implementacijom Chi-2, RFE i IV metoda za odabir značajki, su pokazali najbolje rezultate. Niti jedan od navedena tri modela se nije isticao kao najbolji, ali možemo istaknuti da se model treniran nad skupom podataka proizišlim FIDT tehnikom pokazao najlošijim u usporedbi sa ostalim tehnikama. U radu [4], Stacked Classifier je postigao najbolje rezultate u usporedbi s pojedinačnim modelima. Ovaj klasifikator je dominirao u skoro svim metrikama na sva tri skupa podataka. Jedine iznimke gdje nije dominirao su bile u evaluacijskim metrikama Accuracy kod australskog skupa i F1 kod tajvanskog skupa podataka.

Rad	Trenirani skup podataka	Najbolji model	Accuracy	F1	AUC
[2]	Thera Bank (bez FS)	RF	99,20 %	99,56 %	N/A
[2]	Thera Bank (Hybrid FS)	RF	99,00 %	99,45 %	N/A
[3]	Lending Club (Chi-2)	XGB	77,70 %	78,00 %	0,780
[3]	Lending Club (RFE)	XGB	78,00 %	77,30 %	0,776
[3]	Lending Club (FIDT)	XGB	70,80 %	70,80 %	0,736
[3]	Lending Club (IV)	XGB	77,90 %	77,70 %	0,780
[4]	Australian (IG)	Stacked	86,23 %	84,58 %	0,934
[4]	German (IG)	Stacked	82,80 %	86,35 %	0,944
[4]	Taiwan (IG)	Stacked	85,80 %	51,35 %	0,870

Tablica 4 Opis rezultata evaluacije modela u odabranim publikacijama

2. Odabir značajki

2.1. Što je odabir značajki?

Odabir značajki je ključan proces u pretprocesiranju podataka prije faze modeliranja. Obuhvaća širok spektar algoritama i metoda koji se koriste, između ostalog, za stvaranje reprezentativnijih skupova podataka filtriranjem šumovitih (engl. *noisy*), irelevantnih ili redundantnih uzoraka, što dovodi do optimiziranog treniranja modela strojnog učenja [5].

Metoda odabira značajki ima za cilj odabrati minimalno potreban skup značajki koje najbolje predstavljaju određeni skup podataka, pritom birajući one značajke koje najviše doprinose procjeni vrijednosti predviđane varijable u specifičnom području predikcije. Selekcijom relevantnih značajki kao ulaznih parametara modela za treniranje smanjuje se njegova složenost. Ovisno o dostupnosti oznaka, odabir značajki klasificira se kao nadziran, nenadziran i polunadziran. Većina algoritama za odabir značajki radi s označenim ili neoznačenim podacima. Međutim, ujedinjeni algoritmi za odabir značajki mogu raditi s oba tipa podataka [6].

2.2. Prednosti i nedostatci

Prednosti odabira značajki obuhvaćaju nekoliko ključnih aspekata:

- Bolje razumijevanje strukture podataka:
    - Odabir značajki ključan je za razumijevanje strukture podataka prije izgradnje modela strojnog učenja; potiče jednostavnost, poboljšava performanse modela, pomaže u tumačenju i omogućuje učinkovitiju i efikasniju analizu složenih skupova podataka.
- Poboljšanje performansi modela:
    - Uklanjanjem nepotrebnih ili redundantnih značajki, smanjuje se "šum" u podacima, što omogućuje modelu da bolje generalizira.
- Povećanje interpretabilnosti:

    - Korištenjem manjeg broja značajki, modeli postaju jednostavniji za tumačenje te se lakše prepoznaju uzorci i odnosi unutar podataka, što pomaže prilikom interpretacije rezultata modela.

- Ubrzanje procesa treniranja:
    - Smanjenje vremena treniranja algoritma strojnog učenja, uz smanjenje složenosti algoritma i ubrzavanje cijelog procesa [7].
- Smanjenje troškova prikupljanja podataka:
    - Fokusiranjem na ključne značajke, mogu se optimizirati resursi i izbjeći nepotrebno prikupljanje podataka koji ne doprinose značajno performansama modela, čime se štedi vrijeme i novac.

Nedostaci koji se mogu pojaviti prilikom odabira značajki uglavnom su povezani s nepravilnim korištenjem metoda ili neadekvatnim postavljanjem pragova:

- Gubitak korisnih informacija:
    - Postoji mogućnost da se nepravilnim korištenjem uklone bitne značajke koje smanjuju sposobnost modela.
- Poteškoće u odabiru metode:
    - Različite metode odabira značajki mogu dati različite rezultate, što može otežati odluku o tome koje značajke zadržati. Važno je pravilno odabrati metodu koja je prilagođena tipu podataka.
- Ovisnost o podacima:
    - Odabir značajki temelji se na specifičnom skupu podataka za treniranje, što može rezultirati značajkama koje se ne generaliziraju dobro na nove, neviđene podatke.

2.3. Nadzirani odabir značajki

Ovo je područje najšire primjene metoda za odabir varijabli. U ovom području metode koriste informacije iz oznaka značajki kako bi odabrale relevantne značajke za predikciju ciljne varijable. Metode za odabir označenih podataka mogu se podijeliti u tri glavne kategorije: filter, wrapper i embedded metode.

2.3.1. Filter metode

Filter metode su jednostavne, skalabilne i robusne metode za odabir značajki. Varijable se odabiru na temelju statističkih mjera. Ne odabiru značajke izravno, već rangiraju cijeli skup značajki koristeći evaluacijsku funkciju, tj. evaluacijske kriterije. Odabir značajki vrši korisnik uzimajući u obzir rangiranja (ocjene relevantnosti). Evaluacijska funkcija može se temeljiti na udaljenosti, informacijama (npr. entropiji), točnosti, korelaciji i konzistentnosti. Filter metode ne koriste klasifikacijske modele, već statističke i matematičke funkcije kako bi procijenile kvalitetu značajki [6]. S obzirom da djeluju neovisno o korištenom modelu, zahtijevaju manje računalnog vremena. Ovaj tip metoda koristi različite statističke mjere za procjenu značajki.

Filter metode se mogu podijeliti u dvije glavne kategorije: univarijatne i multivarijatne metode. Univarijatne filter metode procjenjuju svaku značajku pojedinačno, promatrajući njezinu povezanost s ciljnim varijablama, ne uzimajući u obzir međusobnu povezanost značajki. Univarijantne mjere mogu uključivati ANOVA F-test za procjenu varijabilnosti između značajki i ciljne varijable, Mutual Information za mjerenje količine zajedničkih informacija između značajke i ciljne varijable, Chi-Square test za mjerenje povezanosti između kategorijskih značajki i ciljne varijable, te mjere poput Variance Threshold za uklanjanje značajki s vrlo niskom varijancom. Multivarijatne filter metode uzimaju u obzir međusobne odnose između značajki prilikom odabira najvažnijih značajki. One analiziraju više značajki istovremeno, kako bi se procijenila njihova kombinirana povezanost s ciljnim varijablama. Primjer jedne takve metode je korištenje matrice korelacije, koja mjeri stupanj linearne povezanosti između značajki i ciljne varijable, ali i međusobne povezanosti značajki. Značajke koje imaju vrlo visoku međusobnu korelaciju mogu se smatrati redundantnima te se jedna od njih može isključiti kako bi se izbjegla prekomjerna kolinearnost u modelu. Multivarijatne metode su sporije i manje skalabilne od univarijantnih metoda, ali mogu pružiti dublje uvide u odnosima među značajkama.

2.3.2. Wrapper metode

Za razliku od metoda filtera koje ne uključuju treniranje modela strojnog učenja u procesu odabira značajki, metode wrappera uključuju razvoj modela strojnog učenja u procesu odabira značajki kako bi se odabrao najbolji podskup koji vodi do željene izvedbe modela [8]. Drugim riječima, WFS bira značajke na temelju rezultata klasifikatora [9]. Wrapper

metode iterativno isprobavaju različite kombinacije značajki i procjenjuju njihovu izvedbu koristeći određeni model. Ove metode uzimaju u obzir interakcije među značajkama i mogu pronaći optimalni skup značajki specifičan za model. Iako su preciznije, wrapper metode su računalno zahtjevnije, jer zahtijevaju treniranje modela više puta kako bi se evaluirale različite kombinacije značajki. Zbog toga su prikladnije za manje skupove podataka ili kada je dostupna veća računalna snaga.

Kao primjer wrapper metoda, mogu se navesti metode odabira značajki temeljene na pretraživanju koje uključuju pretraživanje unazad (engl. *Backward Elimination*), pretraživanje unaprijed (engl. *Forward Selection*) i rekurzivno pretraživanje značajki (engl. *Recursive Feature Elimination*). Svaka od ovih tehnika koristi model za evaluaciju različitih skupova značajki i odabir onih koje najbolje poboljšavaju performanse modela.

2.3.3. Embedded metode

Embedded metode su prilično slične wrapper metodama, s jedinom razlikom da se odabir značajki vrši samo tijekom treniranja modela. Pomaže u odabiru boljih značajki za taj model u kraćem vremenu. Jedini nedostatak ugrađenih metoda je taj što se odabir značajki odvija u skladu s hipotezom klasifikatora. Međutim, očigledno je da to možda neće dati najbolje rezultate ako se koristi neki drugi klasifikator, a također je ovisno o skupu podataka.

Embedded metode su prilično slične wrapper metodama, s jedinom razlikom da se odabir značajki vrši samo tijekom treniranja modela [10]. Primjer takvih metoda uključuje regularizaciju, kao što su Lasso (L1) i Ridge (L2) regresije, koje penaliziraju složenost modela i automatski smanjuju koeficijente nevažnih značajki na nulu. Embedded metode pomažu u odabiru boljih značajki za određeni model u kraćem vremenu. Jedini nedostatak ovih metoda je taj što se odabir značajki odvija u skladu s hipotezom klasifikatora [10]. Dodatni primjeri embedded metoda uključuju određene algoritme stabala odluke, kao što su RF Feature Importance i GBM Feature Importance, koji automatski odabiru značajke temeljem njihove važnosti za predikciju.

2.4. Nenadzirani odabir značajki

Područje nenadziranog odabira značajki se odnosi na metode koje biraju značajke bez upotrebe oznaka ili ciljanih varijabli, oslanjajući se isključivo na strukturu i odnose unutar samih podataka. Za razliku od nadziranih metoda, koje koriste informacije o klasama ili

vrijednostima ciljanih varijabli za procjenu značajnosti značajki, tehnike nenadziranog učenja izvlače značajne obrasce iz sirovih podataka.

Jedan od češće korištenih pristupa u nenadziranom odabiru značajki je analiza glavnih komponenata (engl. *Principal Component Analysis*, skraćeno PCA). PCA reducira dimenzionalnost podataka identificirajući ortogonalne osi koje maksimiziraju varijabilnost u podacima, čime omogućava odabir značajki koje najviše pridonose ukupnoj varijaciji u skupu podataka. Drugi pristupi uključuju grupiranje (engl. *Clustering*) koji nastoje grupirati slične podatkovne točke zajedno, dok istovremeno održavaju razdvojenost između različitih klastera [11]. Posebno su korisne kada oznake nisu dostupne, ili kada je cilj istražiti strukturu podataka prije primjene nadziranih metoda. One omogućuju otkrivanje latentnih obrazaca i mogu poslužiti kao prvi korak u pripremi podataka za daljnju analizu. Ipak, bez informacija o cilju, postoji rizik da odabrane značajke možda nisu optimalne za specifične zadatke predviđanja, pa se često kombiniraju s nadziranim metodama kako bi se postigli bolji rezultati.

2.5. Polunadzirani odabir značajki

Osnovni koncept polunadziranog učenja je poboljšati sposobnost generalizacije i učenja modela integriranjem označenih i neoznačenih podataka, koristeći informacije o distribuciji, sličnosti oznaka ili drugim strukturalnim značajkama prisutnim u neoznačenim podacima. Polunadzirano učenje poboljšava model koristeći značajnu količinu neoznačenih podataka kako bi uhvatilo latentne uzorke i odnose unutar podataka, čime se povećava njegova robusnost i pouzdanost [12].

Polunadzirane metode su korisne u situacijama kada je dostupno samo ograničen broj oznaka, što je često slučaj u stvarnim aplikacijama gdje ručno označavanje podataka može biti skupo ili vremenski zahtjevno. Ove metode koriste označene podatke za usmjeravanje procesa odabira značajki, dok neoznačeni podaci pomažu u bolje razumijevanju strukture podataka i u identificiranju značajki koje su konzistentne kroz cijeli skup podataka.

Jedan od pristupa polunadziranom odabiru značajki je polunadzirano učenje reprezentacija, gdje se prvo uči reprezentacija podataka iz neoznačenih podataka. Reprezentacija podataka odnosi se na način na koji su podaci strukturirani ili prikazani tako da ističu najvažnije karakteristike ili obrasce unutar podataka. Nakon što model nauči ovu reprezentaciju, koriste se označeni podaci kako bi se odabrale one značajke koje su najrelevantnije za određeni

zadatak. Na taj način, model prvo razumije osnovne strukture u podacima, a zatim koristi oznake za fokusiranje na one značajke koje su najvažnije za predviđanje. Drugi pristup uključuje iterativne metode koje naizmjenično koriste označene i neoznačene podatke kako bi rafinirale skup odabranih značajki.

Prednost polunadziranog odabira značajki je u tome što omogućuje bolju generalizaciju modela, jer koristi informacije iz cijelog skupa podataka, a ne samo iz malog podskupa s oznakama. Međutim, ove metode mogu biti složenije za implementaciju i zahtijevaju ravnotežu između iskorištavanja informacija iz označenih i neoznačenih podataka.

3. Tehnički pregled korištenih metoda i modela

3.1. Tehnike pretprocesiranja podataka

3.1.1. Label Encoding

Label Encoding je tehnika koja se koristi u strojnome učenju za pretvaranje kategorijskih varijabli u numerički oblik. Ova metoda omogućuje modelima strojnog učenja, koji zahtijevaju numeričke podatke, da obrađuju podatke koji su izvorno bili u tekstualnom ili kategorijskom obliku. Tehnika Label Encoding posebno je prikladna za varijable koje imaju ordinalne vrijednosti ili binarnu prirodu. Kada znamo sve moguće vrijednosti kategorijske varijable, njihove kodirane ekvivalente određujemo proizvoljnim odabirom cjelobrojnih vrijednosti koje ćemo im dodijeliti [13]. Ordinalnim varijablama nazivamo sve značajke čije vrijednosti imaju prirodni redoslijed, gdje brojevi zadržavaju semantički smisao tog redoslijeda. Iako je Label Encoding jednostavan i učinkovit za ordinalne i binarne varijable, može uzrokovati probleme kod nominalnih varijabli, gdje se uvodi nepostojeći redoslijed u skup podataka za treniranje modela.

3.1.2. Target Encoding

Target Encoding koristi prosječnu vrijednost ciljne varijable koja odgovara kategorijskim varijablama umjesto samih kategorijskih varijabli [14]. Ova metoda posebno je korisna kada su u pitanju nominalne varijable koje nemaju prirodan poredak. U klasifikacijskim zadacima, to može biti prosječna vjerojatnost pojave ciljne klase, dok u regresijskim zadacima predstavlja prosječnu vrijednost numeričke ciljne varijable.

Prednost Target Encoding-a je u tome što omogućava efikasnije rukovanje varijablama s mnogo kategorija, ne povećavajući dimenzionalnost podataka za razliku od One-Hot Encoding tehnike, te može poboljšati performanse modela jer koristi informaciju iz ciljne varijable. Međutim, postoji rizik od prekomjernog prilagođavanja (engl. *overfitting*), budući da se oslanja na informaciju iz ciljne varijable, što može dovesti do lošije generalizacije na novim podacima.

Kako bi se smanjio rizik od prekomjernog prilagođavanja, Target Encoding se primjenjuje na skupovima podataka za treniranje i testiranje, ali koristeći isključivo informacije iz skupa za treniranje za izračun.

3.1.3. SMOTE

Tehnika sintetičkog naduzorkovanja manjinske klase (engl. *Synthetic Minority Oversampling Technique*, skraćeno SMOTE) smatra se najistaknutijom metodom za rješavanje neuravnoteženih podataka. SMOTE metoda generira nove sintetičke obrasce podataka izvođenjem linearne interpolacije između uzoraka manjinske klase i njihovih k-najbližih susjeda [15]. Koristi k-najbliže susjede kako bi odabrao dva uzorka iz manjinske klase i generirao novi uzorak unutar segmenta između njih. Ova tehnika se često koristi u situacijama gdje je modelu teško naučiti obrasce za manjinsku klasu zbog neuravnoteženosti podataka. Cilj je povećati broj uzoraka manjinske klase bez jednostavnog dupliciranja postojećih uzoraka.

3.1.4. StandardScaler

StandardScaler je tehnika standardizacije podataka koja transformira značajke tako da imaju srednju vrijednost (mean) 0 i standardnu devijaciju 1. Ova metoda je korisna kada značajke imaju različite raspone vrijednosti, što može otežati rad modela, osobito onih osjetljivih na skalu podataka, poput linearnih modela, algoritama temeljenih na udaljenosti i algoritama temeljnih na gradijentnim metodama.

Standardizacija transformira vrijednosti značajke koristeći sljedeću formulu:

$$x_{standardized} = \frac{x - \mu}{\sigma}$$

Gdje je x originalna vrijednost značajke, μ srednja vrijednost značajke u skupu podataka, a σ predstavlja standardnu devijaciju varijable.

3.2. Odabir značajki

3.2.1. Pearson Correlation

Pearson korelacija (engl. *Pearson Correlation*) je statistički alat koji mjeri stupanj linearne povezanosti između dvije numeričke varijable. Vrijednost koeficijenta korelacije može se

kretati između -1 i 1, pri čemu je 1 pokazatelj savršene pozitivne korelacije, -1 savršene negativne korelacije, dok 0 označava da ne postoji linearna povezanost između varijabli. Pearson-ov koeficijent korelacije izračunava se pomoću sljedeće formule:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

gdje x_i i y_i predstavljaju pojedinačne vrijednosti varijabli, a $\overline{x}$ i $\overline{y}$ označavaju njihove prosječne vrijednosti. Pearson-ov koeficijent se koristi za ocjenjivanje linearnosti između varijabli, čime omogućava identifikaciju povezanosti između podataka, kao što su npr. visina i težina osobe. Ova metoda je korisna u analizi podataka u različitim disciplinama jer omogućava mjerenje odnosa između dvije varijable, često se koristi za prepoznavanje multikolinearnosti u skupovima podataka.

3.2.2. Mutual Information

Mutual Information je metoda za odabir značajki koja mjeri koliko informacija dvije varijable dijele međusobno. U kontekstu odabira značajki, ona mjeri koliko informacija značajka pruža o ciljnoj varijabli. Mutual Information može prepoznati i linearne i nelinearne odnose između značajki i ciljne varijable.

Za diskretne varijable, formula za Mutual Information je:

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} P_{X,Y}(x,y)\log(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)})$$

gdje je $P_{X,Y}(x,y)$ zajednička distribucija značajke X i ciljne varijable Y, a $P_X(x)$ i $P_Y(y)$ su marginalne distribucije značajki X i Y.

Za kontinuirane varijable, formula se mijenja i umjesto sume koristi integral:

$$MI(X;Y) = \int_X\int_Y P_{X,Y}(x,y)\log(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)})\,dxdy$$

Mutual Information se koristi za mjerenje količine informacija koje možemo dobiti o ciljnoj varijabli promatranjem svake pojedine značajke. Značajke s većim vrijednostima Mutual Information-a pružaju više informacija o ciljnoj varijabli i stoga se smatraju korisnijima za predikciju.

3.3. Modeli strojnog učenja

3.3.1. Random Forest

Random Forest radi tako što konstruira više stabala odlučivanja tijekom treniranja te na kraju daje predikciju pomoću moda (kod klasifikacije) ili srednje vrijednosti (kod regresije) predikcija pojedinih stabala. Svako stablo odlučivanja u šumi gradi se na temelju nasumično odabranog podskupa podataka za treniranje i nasumično odabranog podskupa značajki. Ova nasumičnost pomaže u smanjenju prekomjernog prilagođavanja i poboljšava sposobnost generalizacije modela. Tijekom procesa treniranja, svako se stablo trenira neovisno koristeći tehniku nazvanu "bagging", koja podrazumijeva uzorkovanje podataka za treniranje s ponavljanjem. Ovaj pristup stvara raznolika stabla koja zajedno tvore robustan model sposoban za hvatanje složenih obrazaca u podacima. Kako bi se napravile predikcije, novi podaci prolaze kroz svako pojedinačno stablo u šumi, a predikcije svih stabala se agregiraju kako bi se dobila konačna predikcija. Kod klasifikacijskih zadataka, mod (najčešće pojavljivana klasa) predikcija pojedinih stabala uzima se kao konačna predikcija, dok se kod regresijskih zadataka računa prosječna vrijednost predikcija pojedinih stabala [16].

Random Forest je robustan i sposoban model za rad s velikim količinama podataka i značajki te dobro podnosi nedosljedne podatke i podatke s odstupanjima. Također automatski procjenjuje važnost značajki, što može biti korisno za interpretaciju modela. Glavne prednosti Random Foresta uključuju otpornost na prekomjerno prilagođavanje zbog ansambl pristupa, mogućnost rada s velikim brojem značajki te sposobnost hvatanja složenih odnosa između značajki.

3.3.2. Extreme Gradient Boosting

Extreme Gradient Boosting (skraćeno XGB), je implementacija algoritma stabla odluke temeljenog na gradijentnom spustu, osmišljenog kako bi točno predvidio ciljnu varijablu. Ovaj algoritam kombinira procjene jednostavnijih modela kako bi spriječio prekomjerno prilagođavanje, uvodeći LASSO (skraćeno L1) i Ridge (skraćeno L2) regularizaciju za „penaliziranje“ složenijih modela. Ciljna funkcija sastoji se od odstupanja modela i regularizacijskog člana, pri čemu se točnost predviđanja određuje uzimajući u obzir odstupanje i varijancu. Proces učenja iterativno dodaje nova stabla koja predviđaju rezidualne vrijednosti prethodnih stabala te ih kombinira za konačnu odluku [16].

3.4. Evaluacijske metrike

3.4.1. Confusion matrix

U slučaju problema binarne klasifikacije, matrica zabune (engl. *Confusion Matrix*) ima dimenziju 2x2, gdje se jedna oznaka smatra "pozitivnom", a druga "negativnom". Elementi matrice karakterizirani su na temelju predviđene oznake (pozitivno, negativno) i rezultata usporedbe predviđene oznake s stvarnom klasom (istinita, lažna) [17]:

1. True Positives (TP) – Broj ispravno predviđenih pozitivnih primjera (kada je model točno predvidio pozitivnu klasu).
2. True Negatives (TN) – Broj ispravno predviđenih negativnih primjera (kada je model točno predvidio negativnu klasu).
3. False Positives (FP) – Broj netočno predviđenih pozitivnih primjera (kada je model netočno predvidio pozitivnu klasu, a stvarna klasa je negativna), također poznat kao Type I error.
4. False Negatives (FN) – Broj netočno predviđenih negativnih primjera (kada je model netočno predvidio negativnu klasu, a stvarna klasa je pozitivna), također poznat kao Type II error.

Navedene kategorije služe za izračun i ostalih evaluacijskih metrika kasnije spomenutih u radu. Matrica zabune je posebno korisna kada se radi o neuravnoteženim skupovima podataka, gdje točnost (engl. *accuracy*) može biti varljiva.

3.4.2. Precision

Preciznost je metrika koja mjeri koliko je točno model predvidio pozitivne primjere. Osjetljiva je na omjer između broja pozitivnih i negativnih instanci u skupu podataka. Kako se povećava udio negativnih instanci, preciznost opada [18]. Drugim riječima, od svih instanci koje je model označio kao pozitivne, preciznost odgovara omjeru onih koji su zaista pozitivni. Preciznost se posebno koristi kada je važno minimizirati broj lažno pozitivnih predikcija (situacije u kojima model pogrešno klasificira negativne primjere kao pozitivne). Formula za preciznost je:

$$Precision = \frac{TP}{TP + FP}$$

Visoka preciznost znači da model rijetko griješi prilikom predikcije pozitivnih primjera, što je važno u situacijama kada je ključno izbjeći lažno pozitivne rezultate.

3.4.3. Recall

Odziv ili osjetljivost (engl. *Recall* ili *Sensitivity*) je metrika koja mjeri koliko je model uspješan u hvatanju svih stvarnih pozitivnih primjera. Drugim riječima, osjetljivost odgovara omjeru stvarno pozitivnih primjera koje je model ispravno predvidio u odnosu na ukupan broj stvarnih pozitivnih primjera. Ova metrika je korisna kada je važno minimizirati broj lažno negativnih predikcija (kada model pogrešno klasificira pozitivne primjere kao negativne). Formula za odziv je:

$$Recall = \frac{TP}{TP + FN}$$

Visoka osjetljivost znači da model uspijeva detektirati većinu stvarno pozitivnih instanci, što je važno u situacijama gdje je kritično ne propustiti pozitivne primjere.

3.4.4. Accuracy

Točnost (engl. *Accuracy*) je najjednostavnija i najčešće korištena metrika za evaluaciju klasifikacijskih modela. Točnost predstavlja omjer broja točnih predikcija (bilo pozitivnih ili negativnih) i ukupnog broja predikcija. Izračunava se prema sljedećoj formuli:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Drugim riječima, točnost pokazuje koliki postotak predikcija modela je ispravan. Iako je vrlo intuitivna, točnost može biti varljiva kada se koristi na neuravnoteženim skupovima podataka. Ako je većina podataka iz jedne klase (npr. 95% negativnih primjera i 5% pozitivnih), model može postići visoku točnost jednostavno predviđajući uvijek većinsku klasu, bez stvarnog uspjeha u razlikovanju među klasama.

Jednostavna je za razumijevanje i implementaciju, ali nije idealna za neuravnotežene skupove podataka jer može biti pristrana prema većinskoj klasi.

3.4.5. F1-score

F1-score je metrika koja uzima u obzir i preciznost i osjetljivost, dajući balansiranu ocjenu između ta dva. F1-score je posebno koristan kada su podaci neuravnoteženi, jer pruža bolju sliku o stvarnoj učinkovitosti modela za obje klase.

Formula za F1-score je:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1-score teži ravnoteži između preciznosti i osjetljivosti, pri čemu je rezultat veći kada su i preciznost i osjetljivost visoki. Ova metrika je posebno korisna u situacijama gdje je jednako važno smanjiti i lažno pozitivne i lažno negativne predikcije. Ova metrika je korisna za neuravnotežene skupove podataka te bolje balansira točnost modela između pozitivnih i negativnih primjera.

3.4.6. AUC

AUC (engl. *Area Under the ROC Curve*) je evaluacijska metrika koja se koristi za procjenu performansi klasifikacijskih modela, osobito u binarnim klasifikacijama. Ova metrika daje uvid u sposobnost modela da razlikuje između dviju klasa, obično pozitivne i negativne. Receiver Operating Characteristic (skraćeno ROC) krivulja je grafički prikaz koji pokazuje odnos između True Positive Rate-a (TPR), odnosno osjetljivosti, i False Positive Rate-a (FPR). TPR predstavlja udio stvarno pozitivnih slučajeva koje model točno klasificira, dok FPR predstavlja udio negativnih slučajeva koje je model pogrešno klasificirao kao pozitivne.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Prilikom uspoređivanja performansi različitih modela strojnog učenja koristeći AUC metriku, potrebno je imati referentnu vrijednost kako bi se utvrdilo je li performansa modela prihvatljiva. AUC se kreće od 50% do 100%, pri čemu 50% odgovara nasumičnom pogađanju, a 100% predstavlja savršenu performansu [19]. Ako je AUC manji od 0.5, to bi značilo da model dosljedno daje pogrešne rezultate. Ovo je korisna metrika jer ne ovisi o postavljenom pragu klasifikacije. Osim toga, AUC je robusna u slučajevima neuravnoteženih podataka, gdje jedna klasa značajno prevladava nad drugom.

4. Metodologija i rezultati

U praktičnom dijelu ovog rada, implementacija svih tehnika i modela provedena je u okruženju Google Colab, koristeći programski jezik Python. Za implementaciju metoda odabira značajki, pretprocesiranja podataka i treniranja modela korištena je biblioteka scikit-learn (sklearn), koja nudi širok spektar alata za strojno učenje.

4.1. Skupovi podataka

Korištena su tri različita skupa podataka koji sadrže informacije o odobrenim kreditima, pri čemu svaka instanca predstavlja pojedinačan kredit s pridruženim podacima o klijentu, transakcijama i ostalim relevantnim informacijama. Iz prvih dvaju skupova podataka izdvojeni su samo podaci dostupni prije odobrenja kredita (engl. *pre-issuance data*) kako bi se mogao kreirati model za donošenje odluke o odobravanju ili odbijanju kreditnog zahtjeva. Treći skup podataka koristi se za modeliranje rizika neispunjavanja obveza (engl. *default*) nakon što je kredit odobren.

Skup podataka (Izvor)	Broj podataka	Broj značajki
Lending Club dataset (Kaggle) [20]	887379	74
Lending Club dataset (Kaggle) [21]	396030	27
Taiwan dataset (UCI) [22]	30000	25

Tablica 5 Specifikacije korištenih skupova podataka

Za potrebe modela koji donosi odluku o odobravanju ili odbijanju kredita, u prvom skupu podataka korištene su isključivo pre-issuance značajke. Prediktivna varijabla u ovom skupu podataka je loan_status, koja u izvornom obliku sadrži nekoliko različitih vrijednosti:

- Current - 601779 uzoraka

- Fully Paid - 207723 uzoraka
- Charged Off - 45248 uzoraka
- Late (31-120 days) - 11591 uzoraka
- Issued - 8460 uzoraka
- In Grace Period - 6253 uzoraka
- Late (16-30 days) – 2357 uzoraka
- Does not meet the credit policy Status: Fully Paid – 1.998 uzoraka
- Default – 1219 uzoraka
- Does not meet the credit policy Status: Charged Off – 761 uzorak

S obzirom na cilj predikcije, krediti sa statusima Current, In Grace Period i Issued su odbačeni, dok su preostali statusi mapirani u dvije kategorije:

- 0 (označava dobar kredit) - uključuje statuse Fully Paid i Does not meet the credit policy. Status: Fully Paid
- 1 (označava loš kredit) - uključuje statuse Charged Off, Late (31-120 days), Late (16-30 days), Default i Does not meet the credit policy. Status: Charged Off

U drugom Lending Club skupu podataka, kao i kod prvog skupa, korištene su samo pre-issuance značajke kako bi se izgradio model za prihvaćanje ili odbijanje kreditnih zahtjeva. U ovom skupu podataka zavisna varijabla loan_status na početku sadrži samo Fully Paid i Charged Off statuse pa je podjela bila sljedeća:

- 0 (označava dobar kredit) - uključuje status Fully Paid
- 1 (označava loš kredit) - uključuje status Charged Off

Treći korišteni skup podataka je tajvanski skup podataka sa UCI repozitorija. Ovaj skup podataka je dosta manji za razliku od prethodna dva, sadrži 30.000 podataka i 25 značajki te se koristi za izgradnju modela za predviđanje neispunjavanja obveza plaćanja već odobrenih kredita na temelju raznih transakcija. Prediktivna varijabla u ovom skupu podataka je default_next_month, koja je binarnog tipa, gdje:

- 0 označava da je klijent izvršio obvezu plaćanja u određenom mjesecu
- 1 označava da klijent nije izvršio svoju obvezu plaćanja u određenom mjesecu

Svi korišteni skupovi podataka u ovom radu su neuravnoteženi u kontekstu omjera binarnih klasa. U svim skupovima podataka prevladava klasa koja predstavlja dobre kredite, tj. klijente s plaćenim obvezama u mjesecu.

Lending Club skupovi podataka sadrže nedostajuće vrijednosti, što je zahtijevalo dodatne postupke za njihovo rješavanje prilikom pripreme podataka. S druge strane, UCI skup podataka ne sadrži nedostajuće vrijednosti, što je olakšalo obradu podataka u ovom slučaju. Također, početne značajke u Lending Club skupovima podataka uključuju i kategoričke i numeričke vrijednosti, dok UCI skup podataka sadrži isključivo numeričke vrijednosti. Ove razlike u strukturi podataka bile su važne pri odabiru metoda za obradu podataka i modeliranje.

4.2. Pretprocesiranje podataka

4.2.1. Čišćenje podataka

Prvi korak pri obradi podataka prvog skupa podataka je bio uklanjanje svih instanci povezanih sa trenutnim statusom, odnosno kredita u trajanju, koji nemaju zabilježen završni ishod. Reduciranjem podataka na samo završene kredite uklonjeno je oko 500 tisuća podataka.

Analizom opisa i vrijednosti svih značajki skupa podataka identificirane su tzv. post-issuance značajke, odnosno značajke koje su proizišle nakon izdavanja kredita, radi prevencije „curenja podataka“ (engl. *Data Leakage*) u modelu predviđanja kreditnog rizika. U tom procesu uklonjeno je 18 značajki. Nakon toga uklonjene su i varijable čije vrijednosti predstavljaju jedinstvene identifikatore klijenata, id i member_id jer ne pridonose predikciji modela. Uklonjena je i značajka url čije su vrijednosti poveznice koje nisu relevantne za ovaj slučaj modeliranja, te je eliminirana i policy_code varijabla iz razloga što sadrži identične vrijednosti za svaki podatak i samim tim nema varijacije. Stupac title je također uklonjen jer predstavlja detaljan opis namjene kredita, a vrijednost namjene kredita se već nalazi unutar skupa podataka pod varijablom purpose. Iz značajke earliest_cr_line, koja predstavlja datum podizanja klijentovog prvog kredita, procesuirana je samo godina za daljnju analizu.

Sljedeći korak je bio eliminacija svih varijabli koje sadrže više od 50% nedostajućih vrijednosti. Nakon ove redukcije, skup podataka je smanjen za 21 značajku. Provedbom ove faze dimenzija skupa podataka je reducirana sa (887379, 74) na (268138, 28).

Što se tiče drugog skupa podataka Lending Club-a, faza čišćenja podataka je bila konciznija za razliku od prethodnog primjera. U ovom skupu podataka nije bilo potrebe za reduciranjem broja podataka na osnovu statusa kredita jer su inicijalni podaci već bili gotovi krediti sa

završenim statusom. Razmatranjem podataka utvrđeno je da su gotovo sve varijable relevantne za model predikcije, osim dviju koje su predstavljale post-issuance tip značajki. Nadalje, stupac title je uklonjen iz istog razloga kao i u prethodnom skupu podataka. Dvije varijable su modificirane metodom ekstrakcije podataka. Značajka address je iskorištena za izdvajanje zip koda, te je na temelju tih podataka kreirana nova varijabla zip_code, dok je originalna varijabla lokacije klijenta uklonjena. Druga značajka, earliest_cr_line, je modificirana na identičan način kao i u prvom skupu.

U tajvanskom skupu podataka uklonjena je značajka ID koja predstavlja jedinstvenu identifikacijsku oznaku podatka i nema doprinosa pri predikciji. Nad preostalim podatcima nije bilo potrebe za čišćenjem jer je ovaj tip modela predikcije koristio post-issuance značajke za razliku od prethodna dva primjera.

4.2.2. Imputacija podataka

Kod Lending Club skupova podataka, preostali stupci s nedostajućim vrijednostima su obrađeni imputacijom na temelju statističkih vrijednosti stupaca. U slučaju numeričkog tipa podataka korištena je srednja vrijednost značajke (engl. *mean*), dok je za kategoričke varijable korištena najčešća vrijednost (engl. *mode*). Tajvanski skup podataka nije sadržao nedostajuće vrijednosti pa nije bilo potrebe za imputacijom podataka.

4.2.3. Enkodiranje ordinalnih značajki

Nakon uklanjanja odstupajućih vrijednosti, u Lending Club skupovima enkodirane su binarne i ordinalne varijable. Binarna varijabla, koja u oba slučaja predstavlja zavisnu varijablu loan_status, enkodirana je u diskretni numerički oblik (0, 1) tehnikom Label Encoding. Ovaj stupac je prvotno sadržavao tekstualne vrijednosti koje su sada mapirane na dvije klase: 0 predstavlja dobar kredit, a 1 loš kredit. Ordinalne varijable su također numerirane korištenjem metode Label Encoding pritom pazeći na redoslijed vrijednosti stupaca. U trećem skupu podataka nije provedena ova tehnika iz razloga što su kategorički podaci već inicijalno bili enkodirani.

```
[239] print(dataset.sub_grade.value_counts().head().to_string())

sub_grade
B3    18650
B4    17699
C1    15880
C2    15339
B2    15134
```

```
[240] # Sortirana lista svih podrazreda
      sub_grade_order = list(dataset.sub_grade.value_counts().sort_index().index)

      # Kreiranje ordinalne mape
      sub_grade_mapping = {grade: idx + 1 for idx, grade in enumerate(sub_grade_order)}

      # Primjena mapiranja / enkodiranje
      dataset['sub_grade'] = dataset['sub_grade'].map(sub_grade_mapping)
```

```
[241] print(dataset.sub_grade.value_counts().head().to_string())

sub_grade
8     18650
9     17699
11    15880
12    15339
7     15134
```

Slika 2 Primjena tehnike Label Encoding na značajci sub_grade

4.2.4. Podjela podataka

Svi skupovi podataka nakon obrade su podijeljeni na skupove za treniranje i testiranje u omjeru 75:25. Ovaj omjer se pokazao optimalnim za treniranje modela jer omogućuje dovoljno podataka za učenje, a pritom osigurava i dostatan postotak za kasniju evaluaciju modela.

4.2.5. Enkodiranje nominalnih značajki

Nakon podjele podataka, korištena je tehnika Target Encoding za enkodiranje nominalnih kategoričkih značajki unutar Lending Club skupova podataka. Odabrana je prosječna vrijednost ciljne varijable. Izračunata je isključivo na nominalnim značajkama iz skupa za treniranje kako bi se spriječio problem „curenja podataka“, odnosno korištenja informacija iz skupa za testiranje u fazu treniranja modela. Konkretno, za svaku nominalnu varijablu izračunat je prosjek ciljne varijable loan_status unutar svake kategorije na temelju podataka iz skupa za treniranje. Nadalje, dobivene vrijednosti su zamijenjene originalnim kategoričkim vrijednostima nakon čega su implementirane i na skupu za testiranje. Ovaj

pristup omogućuje precizniju numeraciju nominalnih varijabli bez povećavanja dimenzionalnosti skupa podataka, što je jedna od ključnih prednosti u odnosu na druge tehnike poput One-Hot Encoding metode. Implementacijom ove metode, skup podataka sada sadrži samo numeričke podatke spremne za skaliranje.

```
[315] print(X_train[nominal_features].head().to_string())
```

```
        home_ownership verification_status             purpose addr_state zip_code
798854       MORTGAGE            Verified  debt_consolidation         CA    925xx
161659            OWN            Verified  debt_consolidation         FL    339xx
37400            RENT        Not Verified                 car         TX    750xx
183956           RENT     Source Verified         credit_card         CA    900xx
328997       MORTGAGE        Not Verified         credit_card         FL    331xx
```

```
[316] train_data = dataset.loc[X_train.index]
      global_mean = train_data['loan_status'].mean()

      for col in nominal_features:
          # Izračunavanje target mean-a na skupu za treniranje
          target_means = train_data.groupby(col)['loan_status'].mean()

          # Primjena target encodinga na skup za treniranje
          X_train[col] = X_train[col].map(target_means)

          # Primjena target encodinga na skup za testiranje koristeći
          # target mean-ove iz skupa za treniranje
          X_test[col] = X_test[col].map(target_means)

          # Zamjena NaN vrijednosti u testnom skupu s globalnim
          # prosjekom ciljne varijable
          X_test[col].fillna(global_mean, inplace=True)
```

```
[317] print(X_train[nominal_features].head().to_string())
```

```
        home_ownership  verification_status   purpose  addr_state  zip_code
798854        0.798443             0.747326  0.764332    0.788984  0.799190
161659        0.762620             0.747326  0.764332    0.748766  0.766004
37400         0.749486             0.825170  0.857795    0.795959  0.807537
183956        0.749486             0.751405  0.800602    0.788984  0.790078
328997        0.798443             0.825170  0.800602    0.748766  0.733657
```

Slika 3 Primjena Target Encoding tehnike

4.2.6. Obrađene značajke

Prije procesa odabira značajki i modeliranja, u sljedećim trima tablicama prikazane su značajke koje su preostale nakon cjelokupnog procesa obrade podataka. Tablice pružaju pregled i opis obrađenih značajki.

Skup podataka	Značajka	Opis
[20]	acc_now_delinq	Broj računa na kojima podnositelj zahtjeva kasni s otplatom.
[20]	addr_state	Savezna država podnositelja zahtjeva.
[20]	collections_12_mths_ex_med	Broj naplata dugovanja u posljednjih 12 mjeseci, isključujući medicinske naplate.
[20]	delinq_2yrs	Broj slučajeva kašnjenja od 30 ili više dana u kreditnoj povijesti podnositelja zahtjeva u posljednje 2 godine.
[20]	inq_last_6mths	Broj kreditnih upita u posljednjih 6 mjeseci (isključujući upite za auto kredite i hipoteke).
[20]	tot_coll_amt	Ukupni iznosi dugovanja koji su ikada bili pod naplatom.
[20]	tot_cur_bal	Ukupan iznos svih računa.
[20]	total_rev_hi_lim	Ukupan visoki revolving kreditni iznos/kreditni limit.
[20]	zip_code	Prve tri znamenke poštanskog broja podnositelja zahtjeva.
[20] [21]	annual_inc	Prijavljeni godišnji prihod podnositelja zahtjeva.
[20] [21]	application_type	Označava je li kredit pojedinačna prijava ili zajednička prijava s dva sudužnika.

[20] [21]	dti	Omjer koji se izračunava koristeći ukupne mjesečne otplate dugova podnositelja zahtjeva na ukupne obveze duga, isključujući hipoteku i traženi LC kredit, podijeljene s podnositeljevim samoprijavljenim mjesečnim prihodom.
[20] [21]	earliest_cr_line	Datum kada je podnositelj zahtjeva prvi put otvorio kreditni račun.
[20] [21]	emp_length	Duljina zaposlenja u godinama. Moguće vrijednosti su između 1 i 10 godina, a vrijednosti manje od 1 ili više od 10 su posebno kategorizirane.
[20] [21]	installment	Mjesečna rata koju duguje korisnik kredita ako kredit bude odobren.
[20] [21]	int_rate	Kamata na kredit.
[20] [21]	grade	Ocjena kredita koju je dodijelio LC (Lending Club).
[20] [21]	home_ownership	Prijavljeni status vlasništva nekretnine koji je podnositelja zahtjeva.
[20] [21]	loan_amnt	Iznos kredita za koji je podnositelj zahtjeva aplicirao.
[20] [21]	open_acc	Broj otvorenih kreditnih računa u kreditnoj povijesti podnositelja zahtjeva.
[20] [21]	pub_rec	Broj negativnih javnih zapisa.

[20] [21]	purpose	Kategorija koju je podnositelj zahtjeva naveo kao svrhu kredita.
[20] [21]	revol_bal	Ukupno stanje revolving kredita.
[20] [21]	revol_util	Stopa iskorištenosti revolving kredita, odnosno iznos kredita koji podnositelj zahtjeva koristi u odnosu na sav dostupni revolving kredit.
[20] [21]	sub_grade	Detaljna ocjena kredita koju je dodijelio LC (Lending Club).
[20] [21]	term	Rok otplate kredita izražen u mjesecima. Može biti 36 ili 60.
[20] [21]	total_acc	Ukupan broj kreditnih računa u kreditnoj povijesti podnositelja zahtjeva.
[20] [21]	verification_status	Označava je li zajednički prihod sudužnika provjerio LC, nije provjeren ili je izvor prihoda provjeren.
[21]	mort_acc	Broj hipotekarnih računa.
[21]	pub_rec_bankruptcies	Broj javnih evidencija o bankrotima.
[21]	zip_code	Poštanski broj podnositelja zahtjeva.
[22]	LIMIT_BAL	Iznos odobrenog kredita u NT dolarima.
[22]	SEX	Spol klijenta.
[22]	EDUCATION	Obrazovanje klijenta.
[22]	MARRIAGE	Bračni status klijenta.

[22]	AGE	Starost u godinama.
[22]	PAY_0	Status otplate u rujnu
[22]	PAY_2	Status otplate u kolovozu
[22]	PAY_3	Status otplate u srpnju
[22]	PAY_4	Status otplate u lipnju
[22]	PAY_5	Status otplate u svibnju
[22]	PAY_6	Status otplate u travnju
[22]	BILL_AMT1	Iznos računa u rujnu
[22]	BILL_AMT2	Iznos računa u kolovozu
[22]	BILL_AMT3	Iznos računa u srpnju
[22]	BILL_AMT4	Iznos računa u lipnju
[22]	BILL_AMT5	Iznos računa u svibnju
[22]	BILL_AMT6	Iznos računa u travnju
[22]	PAY_AMT1	Iznos prethodne uplate u rujnu
[22]	PAY_AMT2	Iznos prethodne uplate u kolovozu
[22]	PAY_AMT3	Iznos prethodne uplate u srpnju
[22]	PAY_AMT4	Iznos prethodne uplate u lipnju
[22]	PAY_AMT5	Iznos prethodne uplate u svibnju
[22]	PAY_AMT6	Iznos prethodne uplate u travnju

Tablica 6 Opisi obrađenih značajki skupova podataka

4.2.7. Odabir značajki

Testirani su rezultati za sve skupove podataka koristeći metode za odabir značajki koje su reducirale skup podataka za 50%, zadržavajući 50% najutjecajnijih značajki. Sljedeće metode odabira značajki su primijenjene:

- PC (engl. *Pearson Correlation*) korištena je za rješavanje multikolinearnosti, tj. visoke korelacije između nezavisnih varijabli. Postavljen je prag korelacije na 0,8, pri čemu je iz svakog para visoko koreliranih značajki izbačena jedna varijabla
- MI (engl. *Mutual Information*) korišten je za mjerenje količine informacije koju svaka značajka dijeli s ciljnom varijablom. Zadržano je 50% najznačajnijih značajki, na temelju praga zadržavanja onih koje pružaju najviše informacija o ciljnoj varijabli.
- RFFI (engl. *Random Forest Feature Importance*) implementiran je kao ugrađena metoda odabira značajki, pri čemu je zadržano 50% najvažnijih značajki na temelju njihovog doprinosa ukupnoj performansi modela.
- XGBFI (engl. *XGB Feature Importance*) također je korišten za rangiranje značajki prema važnosti. Na isti način, zadržano je 50% najvažnijih značajki na temelju njihovog utjecaja na model.

S obzirom kako je skup podataka nakon pretprocesiranja u numeričkom obliku, odnosno sadrži diskretne i kontinuirane numeričke vrijednost, samo su određene filter metode za odabir značajki prikladne za implementaciju. Univarijantna filter metoda MI je pogodna jer se može koristiti za diskretne i kontinuirane vrijednosti, kao i multivarijantna filter metoda PC. Odabrane embedded metode RFFI i XGBFI dodatno rangiraju značajke na temelju njihove važnosti u modelu, čime se osigurava odabir onih najrelevantnijih. Wrapper metode poput RFE nisu korištene zbog svoje računalne zahtjevnosti, budući da zahtijevaju višestruko treniranje modela tijekom procesa odabira značajki, što bi značajno povećalo vrijeme obrade podataka.

4.2.8. Balansiranje podataka

Nakon provedbe metoda za odabir značajki, prije standardizacije podataka korištena je jedna od tehnika prekomjernog uzorkovanja (engl. *oversampling*), to jest SMOTE metoda za generiranje sintetičkih podataka manjinske klase, odnosno loan_status s vrijednošću 1 u slučaju prva dva skupa i default_next_month s vrijednošću 1 u slučaju trećeg skupa, kako bi se izjednačio broj uzoraka s većinskom klasom. Realizacijom ove metode balansiran je skup podataka za treniranje, čime se sprječava da model bude pristran prema većinskoj klasi.

```
[98] from imblearn.over_sampling import SMOTE

     smote = SMOTE(random_state=42)
     X_train, y_train = smote.fit_resample(X_train, y_train)
```

Slika 4 Sintetičko generiranje podataka manjinske klase

4.2.9. Standardizacija podataka

Zadnji korak faze pretprocesiranja bila je standardizacija podataka s ciljem dovođenja svih ulaznih podataka na usporedivu skalu. Upotrijebljena je metoda StandardScaler. Ovim postupkom izbjegnuti su problemi s različitim mjerilima vrijednosti između značajki. Sva tri skupa podataka su skalirana na isti način.

4.3. Modeliranje i evaluacija

Poslije skaliranja relevantnih skupova podataka za treniranje i testiranje, provedeno je modeliranje kako bi se usporedio utjecaj provedenih metoda za odabir značajki. Za usporedbu je korišten i originalni, nereducirani skup značajki kako bi se procijenila korist metoda za odabir značajki.

Korišteni modeli strojnog učenja su Random Forest i XGB, dva često korištena klasifikacijska modela u ovoj domeni predikcije. Za treniranje i evaluaciju korištena su četiri različita skupa značajki:

- Originalni, nereducirani skup značajki
- Skup značajki odabran hibridnom metodom (Pearson, Mutual Information)
- Skup značajki odabran hibridnom metodom (Pearson, RF)
- Skup značajki odabran hibridnom metodom (Pearson, XGB)

Svaka od metoda za odabir značajki testirana je na način da je reduciran skup podataka za 50%, birajući 50% najutjecajnijih značajki u skladu s korištenom metodom. Na ovaj način analizirana je učinkovitost modela na reduciranim skupovima podataka, čime se procijenio stvarni doprinos metoda odabira značajki u odnosu na nereducirani skup značajki.

Za model XGB korišteni su sljedeći hiperparametri u prva dva skupa podataka: max_depth=6, eta=0.1 i num_boost_round=200. U trećem skupu podataka promijenjen je samo hiperparametar num_boost_round=100 zbog manje količine podataka. Konfiguracija Random Forest modela uključivala je hiperparametar n_estimators=200 u oba Lending Club skupa podataka, dok je u trećem skupu smanjen na n_estimators=100, zbog već navedenog razloga. Na temelju evaluacija modela, uspoređene su performanse modela na reduciranim skupovima značajki i nereduciranom skupu, te su analizirani utjecaji odabranih metoda na Accuracy, F1 i AUC metrikama.

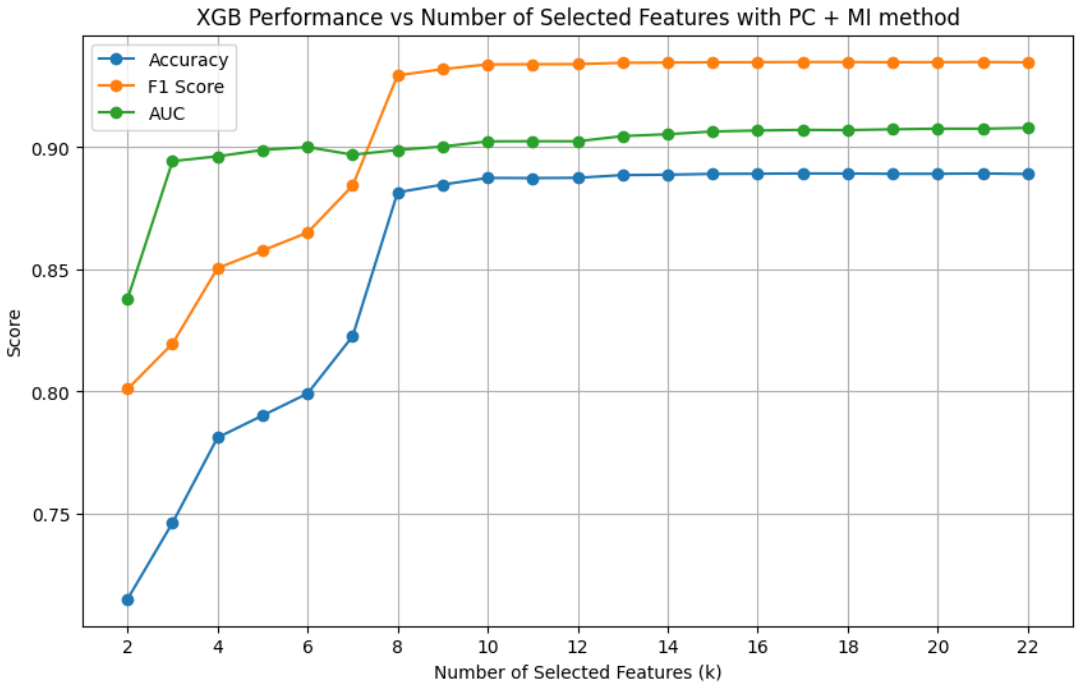
4.4. Rezultati

Skup podataka	FS metoda (broj značajki)	Model	Accuracy	F1	AUC
[20]	PC + MI (14)	RF	77,01%	86,44%	0,6954
[20]	PC + MI (14)	XGB	77,51%	86,86%	0,7115
[20]	PC + RFFI (14)	RF	73,93%	83,66%	0,6851
[20]	PC + XGBFI (14)	XGB	76,74%	86,21%	0,6961
[20]	sve značajke (28)	RF	76,98%	86,41%	0,7062
[20]	sve značajke (28)	XGB	77,61%	86,90%	0,7157
[21]	PC + MI (11)	RF	88,28%	93,06%	0,8920
[21]	PC + MI (11)	XGB	88,86%	93,46%	0,9033
[21]	PC + RFFI (11)	RF	88,07%	92,90%	0,8962
[21]	PC + XGBFI (11)	XGB	88,69%	93,36%	0,9015
[21]	sve značajke (23)	RF	88,79%	93,39%	0,9029

[21]	sve značajke (23)	XGB	88,90%	93,48%	0,9079
[22]	PC + MI (11)	RF	77,07%	46,91%	0,7283
[22]	PC + MI (11)	XGB	76,37%	49,83%	0,7400
[22]	PC + RFFI (11)	RF	78,09%	47,73%	0,7306
[22]	PC + XGBFI (11)	XGB	74,56%	47,35%	0,7297
[22]	sve značajke (23)	RF	77,95%	48,57%	0,7414
[22]	sve značajke (23)	XGB	76,73%	48,30%	0,7480

Tablica 7 Rezultati evaluacija nad odabranim skupovima podataka

Prvi skup podataka [20], pokazuje dosljedne rezultate kroz različite modele i metode odabira značajki. U odnosu na početne značajke, najbolji rezultat u ovom skupu postignut je kombinacijom PC i MI metoda, gdje je korišten model XGB. Točnost modela iznosila je 77,51%, F1-score 86,86%, dok je AUC iznosio 0,7115. Ovi rezultati ukazuju na to da su značajke odabrane metodom MI, u kombinaciji s PC, uspješno doprinosile preciznim predikcijama modela, posebno u pozitivnim klasama. XGB je općenito dao bolje rezultate u usporedbi s RF modelom, što znači da bolje iskorištava složenije interakcije među značajkama u ovom skupu podataka. Drugi Lending Club skup podataka [21] pokazuje značajno bolje rezultate u usporedbi s prvim. Najbolji rezultat u reduciranim skupovima postignut je identičnom kombinacijom kao i u prošlom skupu podataka, tj. metodom PC u kombinaciji s MI tehnikom, pri čemu je model XGB postigao točnost od 88,86%, F1-score od 93,46% te AUC od 0,9033. Korištenje svih značajki nije dovelo do značajnog poboljšanja performansi u odnosu na reduciran skup biran PC+MI metodom, drugim riječima postignuti su gotovo isti rezultati kao i na početnom skupu podataka. U tajvanskom skupu podataka [22], od reduciranih skupova najbolji rezultat postignut je metodom PC u kombinaciji s RFFI metodom, gdje je RF model evaluirao s točnošću od 78,09%, F1-score-om od 47,73% te AUC-om od 0,7306. Općenito gledajući, tajvanski skup podataka je postigao najlošije rezultate u usporedbi s Lending Club skupovima podataka, dok je drugi skup podataka dominirao u svim metrikama naspram ostalih skupova.



Slika 5 Graf omjera broja odabranih značajki PC+MI metodom i performansi XGB modela

Na prikazanom grafu prikazani su omjeri broja odabranih značajki i performansi XGB modela korištenjem PC i MI metoda za odabir značajki na primjeru drugog skupa podataka Lending Club [21]. F1-score (narančasta) raste primjetno do 8 značajki, pri čemu vrhunac doseže oko 10 značajki. Nakon toga, ostaje stabilan s vrlo blagim oscilacijama. Accuracy (plava linija) u početku značajno raste kako se broj odabranih značajki povećava, ali nakon 10 odabranih se stabilizira kao i u F1 slučaju. AUC (zelena linija) započinje na visokoj razini, oko 0.89, već s malim brojem značajki (k=3). Kako se broj značajki povećava, AUC blago raste i doseže stabilnu vrijednost oko 0.91 pri k=10. Nakon toga, dolazi do malog poboljšanja, te se krivulja stabilizira, ostajući na istoj razini bez značajnijih promjena. Očigledno je da model pokazuje optimalan balans performansi već s 8-12 značajki. Daljnje dodavanje značajki ne donosi značajno poboljšanje, a model već u tom rasponu postiže vrlo visoke i stabilne rezultate. Najutjecajnije značajke izabrane navedenom hibridnom tehnikom su: loan_amnt, term, int_rate, emp_length, home_ownership, verification_status, purpose, dti, application_type, mort_acc i zip_code.

Ovaj graf pokazuje važnost odabira značajki, gdje je korištenje manjih, relevantnih skupova značajki dovelo do sličnih performansi modela kao kod korištenja svih dostupnih značajki uz značajnu redukciju ulaznih podataka.

Zaključak

Istraživanje se fokusiralo na optimizaciju procesa odabira značajki u strojnom učenju, s posebnim naglaskom na važnost pravilnog izbora ulaznih podataka i razumijevanje njihovog utjecaja na performanse modela. U ovom radu proučavane su značajke koje pomažu u razlikovanju dobrih i loših kredita. Korištenjem različitih hibridnih metoda odabira značajki, uključujući Pearsonovu korelaciju u kombinaciji s Mutual Information filter metodom, te Random Forest i XGB embedded metodama, postignuti su prilično slični rezultati u usporedbi s originalnim, nereduciranim skupovima značajki. Odabir značajki ima ključnu ulogu ne samo u poboljšanju performansi modela, već i u razumijevanju strukture podataka te pojednostavljenju treniranja modela. Pravilno odabrane značajke pomažu u smanjenju složenosti modela, ubrzavanju treniranja te poboljšanju interpretacije rezultata i generalizacije modela. Također, važno je razumjeti prirodu podataka, jer različite metode za odabir značajki nisu prikladne za sve vrste podataka. Primjena neprikladne metode može rezultirati lošijim performansama modela ili gubitkom važnih informacija. Iz tog razloga je potrebno pažljivo odabrati metode koje odgovaraju specifičnom skupu podataka kako bi se maksimizirala efikasnost modela. Poseban doprinos ovog rada je i grafička analiza koja prikazuje kako broj odabranih značajki utječe na performanse modela. Na temelju grafa moguće je identificirati optimalan broj značajki, koji osigurava ravnotežu između smanjenja složenosti modela i zadržavanja visokih vrijednosti evaluacijskih metrika.

Literatura

[1] Carrera-Rivera A., Ochoa W., Larrinaga F., Lasa G. *How-to conduct a systematic literature review: A quick guide for computer science research*, 2022.

[2] Goel N., Singh D.K. *Reinforcement of the Bank Loan Model using the Feature Selection Method of Machine Learning*, 2023.

[3] Jemai J., Zarrad A. *Feature Selection Engineering for Credit Risk Assessment in Retail Banking*, 2023.

[4] Emmanuel I., Sun Y., Wang Z. *A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method*, 2024.

[5] Grisci B.I., Feltes B.C., Poloni J.F., Narloch P.H., Dorn M. *The use of gene expression datasets in feature selection research: 20 years of inherent bias?*, 2023.

[6] Buyukkececi M., Okur M.C. *A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning*, 2023.as

[7] Piri J., Mohapatra P., Dey R., Acharya B., Gerogiannis V.C., Kanavos A. *Literature Review on Hybrid Evolutionary Approaches for Feature Selection*, 2023.

[8] Zhao T., Zheng Y., Wang Z. *Feature selection-based machine learning modeling for distributed model predictive control of nonlinear processes*, 2023.

[9] A. O. Balogun, S. Basri, L. F. Capretz, S. Mahamad, A. A. Imam, M. A. Almomani, V. E. Adeyemo, A. K. Alazzawi, A. O. Bajeh, and G. Kumar, *Software Defect Prediction Using Wrapper Feature Selection Based on Dynamic Re-Ranking Strategy*, 2021.

[10] Mandal M., Singh P.K., Ijaz M.F., Shafi J., Sarkar R. *A Tri-Stage Wrapper-Filter Feature Selection Framework for Disease Classification*, 2021.

[11] AL-Gburi A.F.J., Nazri M.Z.A., Yaakub M.R.B., Alyasseri Z.A.A. *Multi-Objective Unsupervised Feature Selection and Cluster Based on Symbiotic Organism Search*, 2024.

[12] Li G., Yu Z., Yang K., Lin M., Chen C.L.P. *Exploring Feature Selection with Limited Labels: A Comprehensive Survey of Semi-Supervised and Unsupervised Approaches*, 2024.

[13] Hancock J.T., Khoshgoftaar T.M. *Survey on categorical data for neural networks*, 2020.

[14] Pargent, F., Pfisterer, F., Thomas, J., & Bischl, B. *Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features*, 2022.

[15] Elreedy D., Atiya A.F., Kamalov F. *A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning*, 2023.

[16] Al Shiam S.A., Hasan M.M., Pantho M.J., Shochona S.A., Nayeem M.B., Choudhury M.T.H., Nguyen T.N. *Credit Risk Prediction Using Explainable AI*, 2024.

[17] Markoulidakis I., Rallis I., Georgoulas I., Kopsiaftis G., Doulamis A., Doulamis N. *Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem*, 2021.

[18] Lampert T.A., Gançarski P. *The bane of skew: Uncertain ranks and unrepresentative precision*, 2014.

[19] Imani M., Arabnia H.R. *Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis*, 2023.

[20] Jânio Bachmann, *Lending Club First Dataset*, preuzet sa https://www.kaggle.com/datasets/janiobachmann/lending-club-first-dataset/data, 2019.

[21] Sahil N. Bajaj, *Lending Club Loan Data*, preuzet sa https://www.kaggle.com/datasets/sahilnbajaj/lending-club-loan-data, 2024.

[22] I-Cheng Yeh, *Default of Credit Card Clients*, preuzet sa https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients, 2016.

Skraćenice

AUC	*Area Under the ROC Curve*
FIDT	*Feature Importance using Decision Trees*
FS	*Feature Selection*
IV	*Information Value*
LC	*Lending Club*
LR	*Logistic Regression*
MI	*Mutual Information*
PC	*Pearson Correlation*
ROC	*Receiver operating characteristic*
RF	*Random Forest*
RFFI	*Random Forest Feature Importance*
RFE	*Recursive Feature Elimination*
SMOTE	*Synthetic Minority Over-sampling Technique*
UCI	*University of California Irvine*
UFS	*Univariate Feature Selection*
XGB	*Extreme Gradient Boosting*
XGBFI	*Extreme Gradient Boosting Feature Importance*