Postupci dubinske analize podataka u otkrivanju diskriminacije

Kuljiš, Petra

Sveučilište u Rijeci
Fakultet informatike
i digitalnih tehnologija

Sveučilišni diplomski studij Informacijsko-komunikacijski sustavi

Petra Kuljiš

Postupci dubinske analize podataka u otkrivanju diskriminacije

Diplomski rad

Mentor: izv. prof. dr. sc. Marija Brkić Bakarić

Rijeka, rujan 2024.

Sveučilište u Rijeci
Fakultet informatike i digitalnih tehnologija

Rijeka, 3.6.2024.

Zadatak za diplomski rad

Pristupnica: Petra Kuljiš

Naziv diplomskog rada: Postupci dubinske analize podataka u otkrivanju diskriminacije

Naziv diplomskog rada na engleskom jeziku: Application of data mining techniques in detecting discrimination

Sadržaj zadatka:
Cilj rada je demonstrirati kako se korištenjem naprednih tehnika analize podataka mogu identificirati obrasci diskriminacije koja predstavlja ozbiljan društveni problem. U praktičnom dijelu rada će se odabrani skup podataka primijeniti prikladne tehnike poput vizualizacije podataka, deskriptivne analize, regresijske analize, klasifikacije, klasteriranja i detekcije anomalija.

Mentorica
Izv. prof. dr. sc. Marija Brkić Bakarić

Voditeljica za diplomske radove
Doc. dr. sc. Lucia Načinović Prskalo

Zadatak preuzet: 3.6.2024.

(potpis pristupnice)

Radmile Matejčić 2,
51000 Rijeka, Hrvatska

T: +385 (0)51/584-700
E: ured@inf.uniri.hr
www.inf.uniri.hr

OIB: 64218323816
IBAN: HR1524020061400006966

Sažetak

U diplomskom radu se istražuje primjena dubinske analize podataka u otkrivanju rodne diskriminacije koristeći skup podataka “World Data of Gender Inequality Index” s platforme Kaggle.

Rodna diskriminacija predstavlja ozbiljan društveni problem s globalnim posljedicama, a cilj rada je demonstrirati kako korištenjem naprednih tehnika analize podataka možemo identificirati obrasce diskriminacije koji bi inače ostali neprimijećeni.

Metodologija rada uključuje čišćenje i pretprocesiranje podataka, primjenu tehnika poput deskriptivne analize, regresijske analize, klasifikacije, klasteriranja i detekcije anomalija te vizualizaciju podataka.

Ključne riječi: dubinska analiza podataka; rodna diskriminacija; rodna neravnopravnost; skup podataka; regresijska analiza; klasifikacija; klasteriranje; vizualizacija

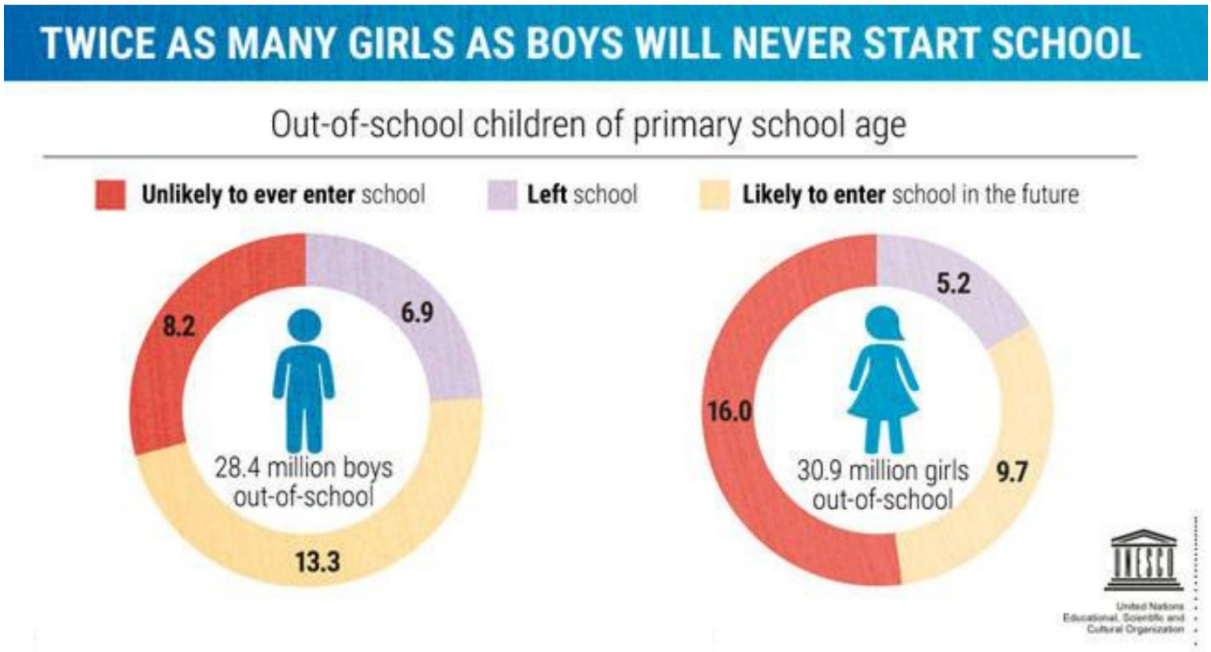
SADRŽAJ

1. Uvod

Riječ diskriminacija potječe od latinske riječi *dis-criminare*, što znači "razlikovati između". U društvenom kontekstu, diskriminacija označava djelovanje temeljem predrasuda koje rezultira nepravednim postupanjem prema ljudima na temelju njihove pripadnosti određenoj kategoriji, bez obzira na njihove individualne zasluge. Ova nepravedna praksa predstavlja ozbiljnu prepreku za postizanje ciljeva jednakosti, razvoja i mira u društvu.

Rodna diskriminacija je oblik diskriminacije koji se odnosi na sustavno nejednako postupanje prema osobama na temelju spola ili roda. Ova nejednakost rezultira privilegiranjem jednog spola ili roda u odnosu na drugi, čime se otežava postizanje ravnoteže i jednakih prilika. Iako je u mnogim društvima postignut napredak u borbi za ravnopravnost spolova, rodna diskriminacija i dalje ostaje duboko ukorijenjena i prisutna širom svijeta. Prema izvještaju UN Women iz 2022. godine, samo 56% zemalja širom svijeta ima zakone koji izričito zabranjuju diskriminaciju temeljem spola na radnom mjestu. Ova statistika ukazuje na pravne nedostatke koji pogoršavaju situaciju.

U obrazovanju, rodna diskriminacija može rezultirati nejednakim prilikama za obrazovanje među djevojčicama i dječacima, što često dovodi do nižih stopa pismenosti i obrazovnih postignuća za žene. Podaci Svjetske banke pokazuju da su žene širom svijeta i dalje u nepovoljnom položaju u pogledu pismenosti u odnosu na muškarce, posebno u manje razvijenim regijama. Na primjer, u nekim regijama poput supsaharske Afrike, razlika između stopa obrazovanja dječaka i djevojčica može biti do 10%, što dovodi do nižih stopa pismenosti i obrazovnih postignuća za žene (slika 1).



Slika 1. Rodne razlike u obrazovanju djevojčica i dječaka Izvor: [1]

U zdravstvu, žene mogu imati ograničen pristup određenim zdravstvenim uslugama i lošije zdravstvene ishode, posebno u područjima poput reproduktivnog zdravlja. Na primjer, u zemljama u razvoju, žene su često izložene većem riziku od smrtnosti zbog komplikacija tijekom poroda.

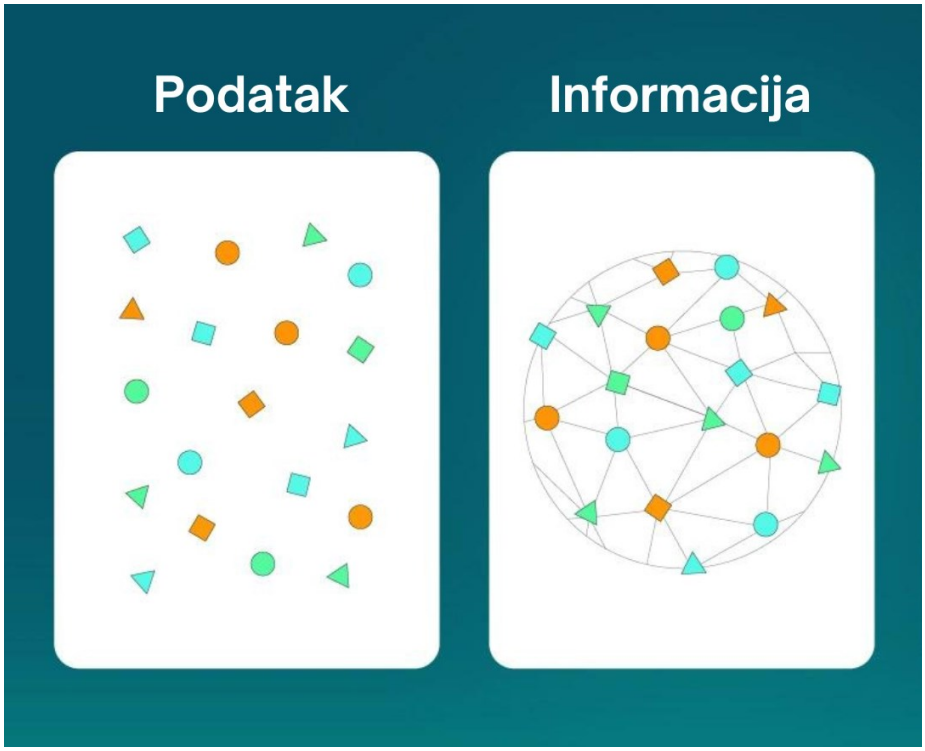
Na tržištu rada, žene često zarađuju manje od muškaraca za iste poslove. Globalni izvještaj o rodnoj razlici iz 2023. godine pokazuje da žene u prosjeku zarađuju oko 16% manje od muškaraca za iste poslove. Ova razlika u plaćama može biti još izraženija u određenim industrijama poput financijskog sektora. Također, prema izvještaju McKinsey & Company, žene čine samo 29% članova upravnih odbora u globalno prepoznatim korporacijama, a još su rjeđe zastupljene na visokim rukovodećim pozicijama.

Još jedan od značajnih problema svakako je politička diskriminacija. Ona se očituje kroz nedovoljnu zastupljenost žena u političkim institucijama i na ključnim pozicijama odlučivanja. Društvene norme i pravni okviri često dodatno pogoršavaju situaciju, jer ne uspijevaju osigurati ravnopravna prava i zaštitu za sve spolove.

Dubinska analiza podataka (engl. *data mining*) nudi moćne alate za razumijevanje i adresiranje složenih problema poput rodne diskriminacije. Ovaj pristup omogućuje istraživačima i donosiocima odluka da bolje razumiju uzroke i manifestacije diskriminacije te da razviju informirane strategije za njezino prevladavanje, pri čemu je važno osigurati etički pristup u svim fazama analize.

2. Dubinska analiza podataka

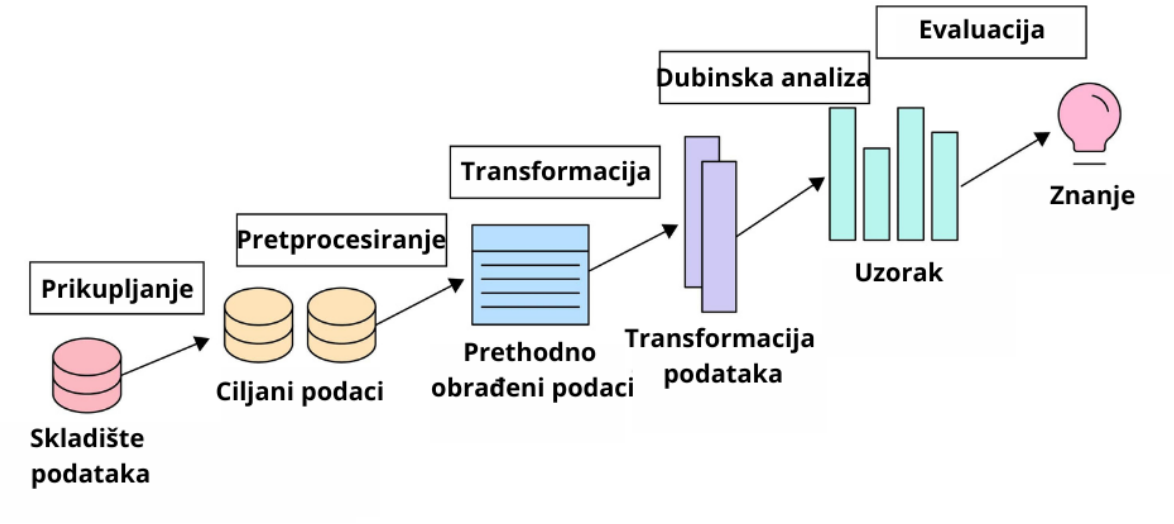
U modernom društvu, obilježenom brzim napretkom tehnologije i sveprisutnom digitalizacijom, pojam podatak postaje sve značajniji. Podaci se mogu definirati kao zabilježene činjenice, brojke ili druge vrste sirovih zapisa koji sami po sebi nemaju značenje. Oni čine osnovu za stvaranje korisnih informacija, no njihovo značenje nije uvijek očigledno. Informacije, s druge strane, predstavljaju organizirane i strukturirane podatke koji su analizirani i interpretirani kako bi se stvorila značenja i obrasci. Dok su podaci jednostavni i često neuredni, informacije su rezultat obrade i analize koji omogućuju razumijevanje i donošenje odluka (slika 2).



Slika 2. Razlika podatka i informacije Izvor: [2]

Proces pretvaranja podataka u informacije zahtjeva sofisticirane metode analize koje omogućuju prepoznavanje značajnih obrazaca i odnosa među podacima. Iako se današnje društvo temelji na informacijama, što je rezultat spajanja računalne tehnologije i komunikacije, većina tih informacija ostaje u svom sirovom obliku, tj. podacima.

Početkom 1990-ih, razvila se ideja o kontinuiranom procesu otkrivanja znanja iz baza podataka (engl. *Knowledge Discovery in Databases*, KDD). Ovaj sveobuhvatni proces omogućuje pretvaranje sirovih podataka u korisne i značajne informacije, tj. znanje [3]. Proces KDD-a uključuje korake od prikupljanja i pripreme podataka, preko primjene analitičkih tehnika, pa sve do interpretacije i prezentacije rezultata (slika 3).



Slika 3. Proces otkrivanja znanja u bazama podataka (KDD) Izvor: [4]

U KDD metodi, četvrta faza je dubinska analiza podataka koja igra ključnu ulogu u izvlačenju korisnih informacija iz podataka. Dubinska analiza podataka je analitička faza prepoznavanja obrazaca i izdvajanje detalja o bazama podataka korištenjem inteligentnih metoda, uključujući strojno učenje, statističku analizu i sustave baze podataka.

Prepoznavanje obrazaca u podacima nije nova pojava, ljudska potreba za traženjem obrazaca prisutna je već tisućama godina. Od najranijih vremena, ljudi nastoje razumjeti obrasce u okolišu kako bi unaprijedili svoje živote, npr. mornari koji su proučavali obrasce kretanja zvijezda na nebu kako bi se orijentirali na moru i predvidjeli vremenske uvjete ili političari i analitičari koji istražuju obrasce u biračkom ponašanju kako bi bolje razumjeli društvene trendove.

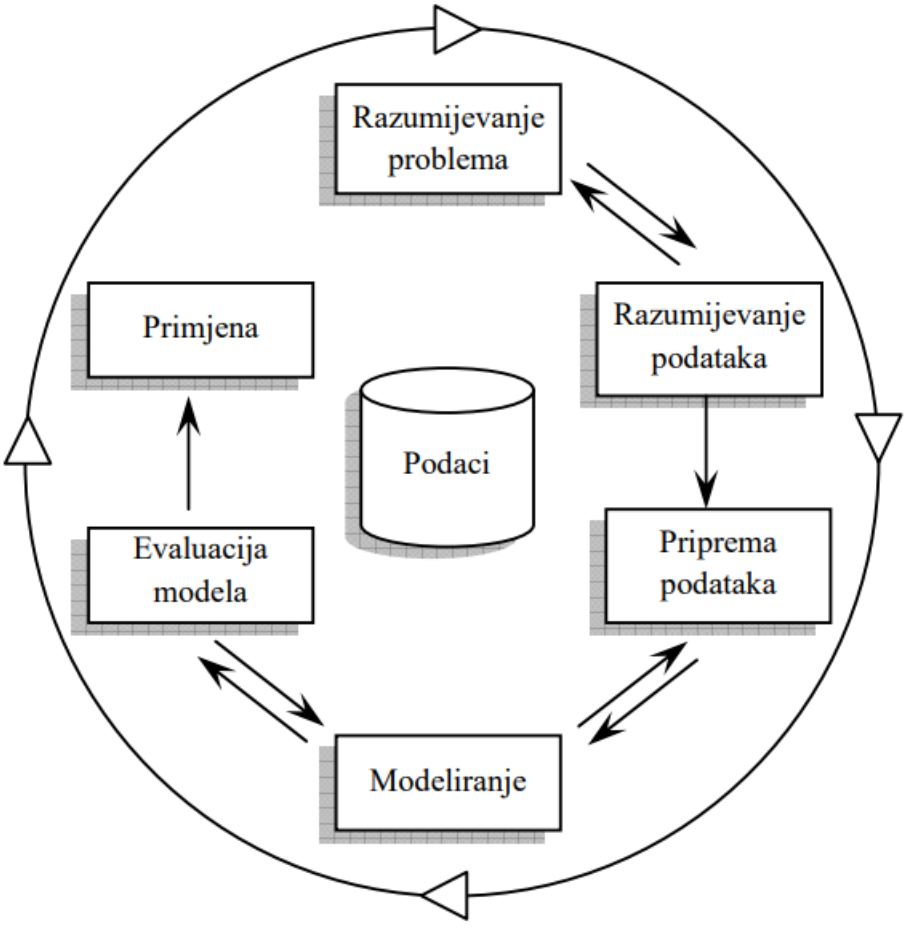
U znanstvenom svijetu posao znanstvenika je identificirati obrasce u prirodnim fenomenima i razviti teorije koje pomažu u predviđanju budućih događaja. Danas, s razvojem tehnologije i dostupnošću velikih količina podataka, dubinska analiza omogućuje otkrivanje skrivenih obrazaca u velikim količinama podataka, pružajući uvide koji mogu unaprijediti poslovne strategije, poboljšati zdravstvenu skrb i razumjeti društvene dinamike.

Metode dubinske analize podataka, uključujući deskriptivnu statistiku, regresijsku analizu, klasifikaciju, klasteriranje, analizu glavnih komponenti (engl. *Principal Component Analysis*, PCA) i vizualizaciju podataka, pružaju uvid u ključne faktore i obrasce unutar velikih skupova podataka. Korištenjem naprednih analitičkih tehnika i algoritama strojnog učenja, dubinska analiza omogućava identifikaciju skrivenih veza koje mogu pomoći u razvoju učinkovitih politika i intervencija za smanjenje rodne diskriminacije i unapređenje ravnoteže spolova.

2.1. Faze dubinske analize podataka

U ranim fazama razvoja dubinske analize podataka, projekti su se često provodili na različite načine, pri čemu je svaki analitičar podataka razvijao vlastiti pristup problemu, često metodom pokušaja i pogrešaka. Kako su se tehnike dubinske analize razvijale, a poslovne potrebe za njima rasle, postojala je sve veća potreba za boljim razumijevanjem i standardizacijom procesa otkrivanja znanja. Ova potreba za standardizacijom bila je motivirana željom da se potencijalnim klijentima pokaže da je dubinska analiza podataka dovoljno zrela za implementaciju kao ključni element poslovnih strategija.

Taj je razvoj doveo do stvaranja industrijskog standarda za proces dubinske analize podataka (engl. *Cross-Industry Standard Process for Data Mining*, CRISP-DM), kojeg su 2000. godine predstavili Chapman i suradnici. CRISP-DM metodologija dizajnirana je da bude neovisna o izboru alata za dubinsku analizu podataka, industrijskom segmentu i specifičnoj primjeni ili problemu koji se rješava. Ova metodologija definira ključne korake u procesu otkrivanja znanja, a premda većina projekata zahtijeva višestruke iteracije pojedinih koraka ili sekvenci koraka, osnovne smjernice CRISP-DM metodologije korisne su za analitičare podataka, ali i za klijente (slika 4).



Slika 4. Faze dubinske analize podataka prema CRISP-DM standardu Izvor: [5]

Ključne faze CRISP-DM metodologije su:

- Razumijevanje problema: definiranje poslovnog cilja i razumijevanje problema koji se treba riješiti
- Razumijevanje podataka: sakupljanje i razumijevanje podataka koji su potrebni za rješavanje problema
- Priprema podataka: korištenje raznih operacija za čišćenje i transformaciju podataka kako bi bili spremni za modeliranje
- Modeliranje podataka: analiziranje dobivenih podataka iz prethodne faze kako bi se otkrilo potrebno znanje
- Evaluacija modela: procjena kvalitete i pouzdanosti dobivenog modela
- Primjena rezultata: integriranje novostečenog znanja

Osnovna prednost CRISP-DM metodologije je njena fleksibilnost i prilagodljivost različitim vrstama podataka i specifičnim problemima. Metodologija se ne odnosi samo na tehničke aspekte analize, već također pomaže u definiranju poslovnih ciljeva i postizanju rezultata koji su primjenjivi u stvarnom svijetu. Ovakav pristup donosi boljem razumijevanju podataka i osigurava kvalitetnije donošenje odluka, što je ključno u kontekstu otkrivanja i rješavanja problema diskriminacije [6, 7].

2.2. Tehnike dubinske analize podataka

Cilj dubinske analize podataka je identificirati nove, potencijalno korisne i razumljive korelacije i obrasce u postojećim podacima. Kako bi izvukli ove relevantne informacije iz skupova podataka, istraživači koriste mnogo naprednih algoritama i tehnika koji se klasificiraju u dvije kategorije: deskriptivne i prediktivne.

Prediktivne tehnike koriste se za prognoziranje budućih događaja na temelju postojećih podataka. Najčešće tehnike uključuju klasifikaciju, regresiju, predikciju vremenskih serija i detektiranje anomalija. Ove metode omogućuju donošenje informiranih odluka na temelju prethodno utvrđenih obrazaca u podacima.

Deskriptivne tehnike usredotočene su na opisivanje osnovnih karakteristika podataka i identifikaciju postojećih obrazaca bez predviđanja budućih vrijednosti. U ovu kategoriju spadaju tehnike kao što su klasteriranje, asocijativna pravila, PCA i komparativna analiza. Ove metode omogućuju bolje razumijevanje strukture i odnosa unutar podataka.

Integracija deskriptivnih i prediktivnih tehnika omogućuje cjelovitu analizu podataka, pri čemu se prvo prepoznaju obrasci i odnosi, a zatim se ti uvidi koriste za predviđanje i donošenje strateških odluka.

Tablica 1. Pregled ključnih tehnika

Kategorija	Tehnike	Opis
Prediktivne	Klasifikacija	Razvrstava podatke u kategorije na temelju značajki (npr. predviđanje diskriminacije pri zapošljavanju na temelju neke karakteristike)
	Regresija	Predviđa kontinuirane vrijednosti (npr. procjena razlika u plaćama između muškaraca i žena na temelju podataka o iskustvu, obrazovanju i industriji)
	Predikcija vremenskih serija	Predviđa buduće vrijednosti na temelju vremenskih podataka (npr. analiza trendova u rodnoj nejednakosti u pristupu obrazovanju tijekom godina)
	Detektiranje anomalija	Identificira neuobičajene podatke koji odstupaju od normi (npr. otkrivanje nerazmjerno visokih stopa odbijanja kandidata određenog spola ili etničke pripadnosti)
Deskriptivne	Klasteriranje	Grupira slične podatke u klastere bez unaprijed definiranih oznaka (npr. segmentacija zaposlenika prema različitim karakteristikama kako bi se otkrili obrasci diskriminacije)
	Asocijativno pravilo	Otkiva veze između varijabli (npr. otkrivanje povezanosti između spola kandidata i učestalosti zapošljavanja u određenim industrijama)
	PCA	Smanjuje dimenzionalnost podataka za vizualizaciju skrivenih obrazaca (npr. analiza faktora koji najviše pridonose rodnoj nejednakosti u plaćama)
	Komparativna analiza	Uspoređuje različite skupove podataka za otkrivanje sličnosti i razlika (npr. usporedba plaća između muškaraca i žena u različitim sektorima)

2.3. Primjena dubinske analize podataka

Sredinom 1990-ih programeri i korisnici sustava za podršku odlučivanju u područjima kao što su financije (npr. aplikacija za odobravanje kredita i otkrivanje prijevara), marketing i analiza prodaje (npr. obrasci kupovine i predviđanje prodaje) počeli su pokazivati veliki entuzijazam u vezi s poslovnim značajem aplikacija za dubinsku analizu podataka. Tijekom slijedećih nekoliko godina međunarodne konferencije, časopisi i knjige sve su češće izvještavali o napretku, alatima i primjenama u drugim područjima kao što su biomedicinska informatika, inženjerstvo, fizika, provedba zakona, poljoprivreda i mnogim drugim.

Danas se dubinska analiza podataka koristi gotovo u svim industrijama i znanstvenim disciplinama, gdje igra ključnu ulogu u donošenju odluka, poboljšanju operativne učinkovitosti i omogućavanju inovacija. U financijama, koristi se za procjenu kreditnog rizika, otkrivanje prijevara i analizu investicijskih strategija. Zdravstvena industrija koristi ove tehnike za dijagnozu bolesti, personalizaciju tretmana, te analizu kliničkih podataka, što doprinosi preciznijoj i učinkovitijoj medicinskoj skrbi.

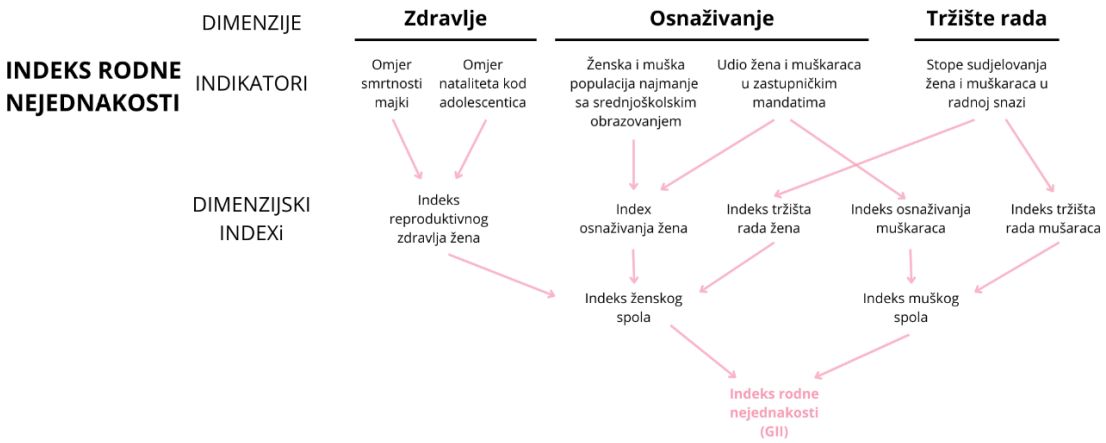
Neke od najaktivnijih primjena dubinske analize podataka nalaze se u području marketinga i prodaje. U tim domenama, tvrtke posjeduju goleme količine precizno zabilježenih podataka, za koje se tek nedavno shvatilo da su iznimno vrijedni. U ovim aplikacijama, naglasak je na preciznim predviđanjima, dok je način na koji se odluke donose često sporedan i manje bitan [8].

Jedno od suvremenih i ključnih područja primjene dubinske analize podataka je i otkrivanje diskriminacije. Ova tehnika omogućava analizu velikih skupova podataka kako bi se identificirali obrasci i anomalije koje upućuju na diskriminaciju po različitim osnovama, poput roda, rase ili dobi. Analizom podataka iz različitih sektora, poput zapošljavanja, obrazovanja i zdravstva, moguće je otkriti nepravedne prakse i razviti strategije za njihovo uklanjanje.

3. Analiza podataka i analitičke tehnike

Za otkrivanje diskriminacije korišteni su podaci iz skupa podataka "WORLD DATA OF Gender Inequality Index" s web stranice Kaggle. Ovi podaci omogućuju analizu rodne nejednakosti među zemljama na temelju različitih čimbenika.

Indeks rodne nejednakosti (engl. *Gender Inequality Index*, GII) odražava rodnu nepravdu u tri dimenzije – reproduktivnom zdravlju, osnaživanju i tržištu rada (slika 5) gdje GII pokazuje koliko rodna nejednakost utječe na ljudski razvoj u različitim zemljama. Vrijednost GII-a se kreće od 0, gdje su žene i muškarci potpuno jednaki, do 1, gdje jedan spol ima lošije rezultate u svim dimenzijama.



Slika 5. Računanje GII-a Izvor: [9]

3.1. Priprema podataka

Skup podataka preuzet je u CSV formatu (slika 6), što omogućuje jednostavno učitavanje i obradu pomoću programskog jezika Python. CSV format praktičan je jer podržava različite metode analize i čišćenja podataka, što je ključno za osiguravanje kvalitete analize. Varijable u skupu podataka obuhvaćaju razne dimenzije rodne nejednakosti, što omogućava detaljno istraživanje utjecaja različitih faktora na rodnu diskriminaciju na globalnoj razini. U nastavku je dan njihov popis.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	HDI rank	Country	HUMAN DI	GII VALUE	GII RANK	Maternal_r	Adolescen	Seats_parl	F_seconda	M_seconda	F_Labour_f	M_Labour_force	
2	1	Switzerlan	VERY HIGH	0.018	3	5	2.2	39.8	96.9	97.5	61.7	72.7	
3	2	Norway	VERY HIGH	0.016	2	2	2.3	45	99.1	99.3	60.3	72	
4	3	Iceland	VERY HIGH	0.043	8	4	5.4	47.6	99.8	99.7	61.7	70.5	
5	4	Hong Kong	VERY HIGH	..	..	..	1.6	..	77.1	83.4	53.5	65.8	
6	5	Australia	VERY HIGH	0.073	19	6	8.1	37.9	94.6	94.4	61.1	70.5	
7	6	Denmark	VERY HIGH	0.013	1	4	1.9	39.7	95.1	95.2	57.7	66.7	
8	7	Sweden	VERY HIGH	0.023	4	4	3.3	47	91.8	92.2	61.7	68	
9	8	Ireland	VERY HIGH	0.074	21	5	5.9	27.3	88.1	86	56.5	68.6	
10	9	Germany	VERY HIGH	0.073	19	7	7.5	34.8	96.1	96.5	56.8	66	
11	10	Netherlan	VERY HIGH	0.025	5	5	2.8	39.1	89.8	92.7	62.4	71.3	

Slika 6. Izgled skupa podataka i varijabli

- HDI Rank - rangiranje zemalja prema indeksu ljudskog razvoja (engl. *Human Development Index*, HDI) (mjeri opću razinu ljudskog razvoja u zemlji na temelju faktora kao što su zdravlje, obrazovanje i standard života), gdje niži brojčani rangovi (npr. 1, 2, 3) označavaju visoki ljudski razvoj, a viši brojčani rangovi (npr. 190, 191) označavaju niži ljudski razvoj
- Country - naziv zemlje
- Human development – razina ljudskog razvoja u zemlji
- GII value - vrijednost GII-ja za zemlju
- GII rank - rangiranje zemalja prema njihovom GII-u
- Maternal_mortality - broj žena umrlih od uzroka povezanih s trudnoćom na 100 000 živorođene djece (pokazuje pristup zdravstvenim uslugama za žene)
- Adolescent_birth_rate - broj rođenja u žena u dobi od 15 do 19 godina na 1000 žena u dobi od 15 do 19 godina (pokazuje obrazovne i zdravstvene uvjete za mlade djevojke)
- Seats_parliamentt(% held by women) - udio mjesta koje drže žene u nacionalnom parlamentu izražen kao postotak od ukupnog broja mjesta (mjeri političku participaciju i osnaživanje žena u političkim strukturama)
- F_secondary_educ - postotak žena u dobi od 25 i više godina koje je steklo srednjoškolsko obrazovanje (pokazuje razinu obrazovanja žena u zemlji)
- M_secondary_educ - postotak muškaraca u dobi od 25 i više godina koje je steklo srednjoškolsko obrazovanje (koristi se za usporedbu s obrazovanjem žena i procjenu ravnoteže spolova u obrazovanju)
- F_Labour_force - postotak žena koje sudjeluju u radnoj snazi (pokazuje ekonomski angažman žena u zemlji)
- M_Labour_force - postotak muškaraca koje sudjeluju u radnoj snazi (koristi se za usporedbu sa sudjelovanjem žena i procjenu rodne ravnoteže na tržištu rada)

Analiza uključuje pripremu podataka, klasifikaciju, primjenu algoritama strojnog učenja, regresijsku analizu, klasteriranje i detekciju anomalija. Cilj je razumjeti čimbenike koji utječu na rodnu nejednakost u različitim zemljama i identificirati obrasce i anomalije.

Priprema podataka ključan je korak u procesu analize jer osigurava da se podaci pravilno transformiraju i očiste što omogućuje preciznije modele i pouzdanije rezultate. Kvalitetna priprema podataka minimizira rizik od pogrešnih interpretacija, što je posebno važno u istraživanju osjetljivih tema kao što je rodna nejednakost.

3.1.1. Učitavanje podataka

Prvi korak u analizi podataka je njihovo učitavanje i priprema za daljnju obradu. Kako bi se omogućio pristup skupu podataka s Kaggle platforme, kreira se direktorij i postavlja datoteka

s API ključem. Ova metoda osigurava sigurno i automatizirano preuzimanje podataka, što pojednostavljuje proces obrade. Skup podataka se zatim učitava u Pandas DataFrame, stvarajući osnovu za daljnju analizu. Pandas DataFrame omogućuje manipulaciju podacima na efikasan način, što je ključno za preciznu identifikaciju obrazaca rodne nejednakosti.

Za bolje razumijevanje sadržaja i kvalitete podataka potrebno je ispitati njihovu strukturu, kvalitetu i osnovne karakteristike, za to je korištena metoda *shape*.

Metoda *info()* daje detaljan pregled skupa podataka, što omogućuje prepoznavanje potencijalnih problema s podacima već u ranoj fazi analize. Na slici 7 prikazani su nazivi stupaca, broj ispravnih vrijednosti i tip podataka.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 13 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   HDI rank                              195 non-null    int64
 1   Country                               195 non-null    object
 2   HUMAN DEVELOPMENT                     195 non-null    object
 3   GII VALUE                             195 non-null    object
 4   GII RANK                              195 non-null    object
 5   Maternal_mortality                    195 non-null    object
 6   Adolescent_birth_rate                 195 non-null    float64
 7   Seats_parliamentt(% held by women)    195 non-null    object
 8   F_secondary_educ                      195 non-null    object
 9   M_secondary_educ                      195 non-null    object
 10  F_Labour_force                        195 non-null    object
 11  M_Labour_force                        195 non-null    object
 12  Unnamed: 12                           0 non-null      float64
dtypes: float64(2), int64(1), object(10)
memory usage: 19.9+ KB
```

Slika 7. Detaljan pregled skupa podataka

3.1.2. Čišćenje podataka

Nakon preliminarne analize podataka, identificirani su slijedeći problemi:

- „Unnamed:12” - stupac 12 sadrži samo NaN vrijednosti i nije koristan
- Nulte vrijednosti su predstavljene s dvije točke („..”) umjesto standardnih NaN vrijednosti
- Neki stupci nemaju točan tip podatka

Kako bismo osigurali da podaci budu kvalitetni za daljnju analizu, ovi problemi moraju biti riješeni. Uklanjanje nepotrebnih stupaca i zamjena nedostajućih vrijednosti osigurava da model koristi samo relevantne i pouzdane podatke, što je ključno za precizno otkrivanje obrazaca diskriminacije.

```
data.drop(columns=['Unnamed: 12'], inplace=True)
data.replace('..', np.nan, inplace=True)
numeric_columns = ['GII VALUE', 'GII RANK', 'Maternal_mortality',
               'Seats_parliamentt(% held by women)','F_secondary_educ',
               'M_secondary_educ', 'F_Labour_force', 'M_Labour_force']
data[numeric_columns] = data[numeric_columns].apply(pd.to_numeric, errors='coerce')
```

Najprije se uklanja nepotrebni stupac „Unnamed:12" koji se u potpunosti sastoji od NaN vrijednosti. Kako Pandas biblioteka ne prepoznaje automatski „.." kao prazne, potrebno ih je zamijeniti s NaN. Zbog prisutnosti nenumeričkih vrijednosti („.."), neki stupci prije te izmjene mogu netočno biti prikazani kao stringovi te je potrebno promijeniti tip podatka u numerički.

Koristeći metodu *isna().sum()* provjerava se broj podataka koji sadrže prazne vrijednosti.

HDI rank	0
Country	0
HUMAN DEVELOPMENT	0
GII VALUE	25
GII RANK	25
Maternal_mortality	14
Adolescent_birth_rate	0
Seats_parliamentt(% held by women)	2
F_secondary_educ	18
M_secondary_educ	18
F_Labour_force	15
M_Labour_force	15

Slika 8. Prikaz broja praznih vrijednosti

Na slici 8 vidimo da ima 25 nepotpunih redaka koje je potrebno ukloniti metodom *dropna()*.

Radi se provjera duplih vrijednosti korištenjem metode *duplicated()*, te se ponovno ispisuje tablica praznih vrijednosti radi zadnje provjere (slika 9). Ovaj korak je važan za osiguranje da kasnije analize i modeli ne budu pogrešno interpretirani zbog nedostajućih podataka.

HDI rank	0
Country	0
HUMAN DEVELOPMENT	0
GII VALUE	0
GII RANK	0
Maternal_mortality	0
Adolescent_birth_rate	0
Seats_parliamentt(% held by women)	0
F_secondary_educ	0
M_secondary_educ	0
F_Labour_force	0
M_Labour_force	0

Slika 9. Prikaz praznih vrijednosti nakon čišćenja

Nakon primjene spomenutih koraka, uređeni skup podataka biti će kvalitetniji i spreman za daljnju analizu. Čišćenje podataka je od fundamentalne važnosti jer nudi jasniju sliku međusobnom djelovanju različitih varijabli i na koji način utječu na rodnu nejednakost.

4. Deskriptivna analiza

Deskriptivna analiza skupa podataka Gender Inequality Index fokusira se na generiranje sažetih statistika i vizualizaciju ključnih metrika koje mogu ukazivati na prisutnost rodne diskriminacije. Ova analiza omogućava dubok uvid u osnovne karakteristike podataka, identificirajući ključne obrasce i trendove koji mogu ukazivati na diskriminaciju. Uspoređujući različite zemlje i njihove GII vrijednosti, analiza omogućava prepoznavanje specifičnih područja u kojima je rodna nejednakost najizraženija.

```
key_metrics = data[['GII VALUE', 'Maternal_mortality',
                'Adolescent_birth_rate',Seats_parliamentt(% held by
                women)', 'F_secondary_educ','M_secondary_educ',
                 'F_Labour_force', 'M_Labour_force']]

key_metrics.describe()
```

Metoda pruža uvid u osnovne statističke mjere za određene varijable, uključujući prosjek, standardnu devijaciju, minimum, maksimum i kvartile (slika 10).

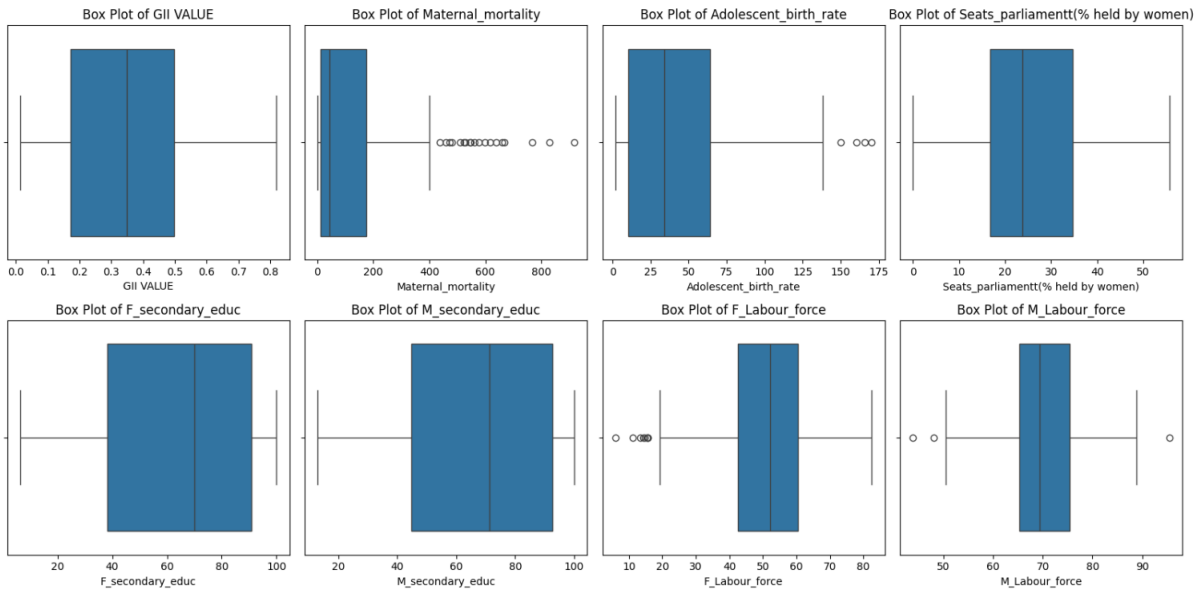
	GII VALUE	Maternal_mortality	Adolescent_birth_rate	Seats_parliamentt(% held by women)	F_secondary_educ	M_secondary_educ	F_Labour_force	M_Labour_force
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	0.339353	139.592814	43.623952	25.297605	62.779641	67.069461	50.218563	70.088024
std	0.195194	195.249768	38.311155	12.398440	29.622536	26.654098	15.351420	8.472413
min	0.013000	2.000000	1.900000	0.000000	6.400000	13.000000	6.000000	43.900000
25%	0.172000	10.500000	10.450000	16.750000	38.100000	44.750000	42.550000	65.300000
50%	0.349000	45.000000	33.900000	23.800000	69.900000	71.200000	52.100000	69.400000
75%	0.498500	175.000000	64.200000	34.650000	90.750000	92.700000	60.350000	75.450000
max	0.820000	917.000000	170.500000	55.700000	100.000000	100.000000	82.500000	95.500000

Slika 10. Sumarne statistike

4.1. Vizualizacija deskriptivne analize

Korištenjem različitih vizualizacija, u ovom slučaju Box Plot-ova i korelacijske matrice, može se dobiti dublji uvid u raspodjelu vrijednosti i odnose između ključnih varijabli u skupu podataka.

Box Plot-ovi su korišteni za prikaz raspodjele na ključne metrike unutar skupa podataka, što omogućava identifikaciju odstupanja, opsega raspona vrijednosti i uvid u centralne tendencije podataka (slika 11).

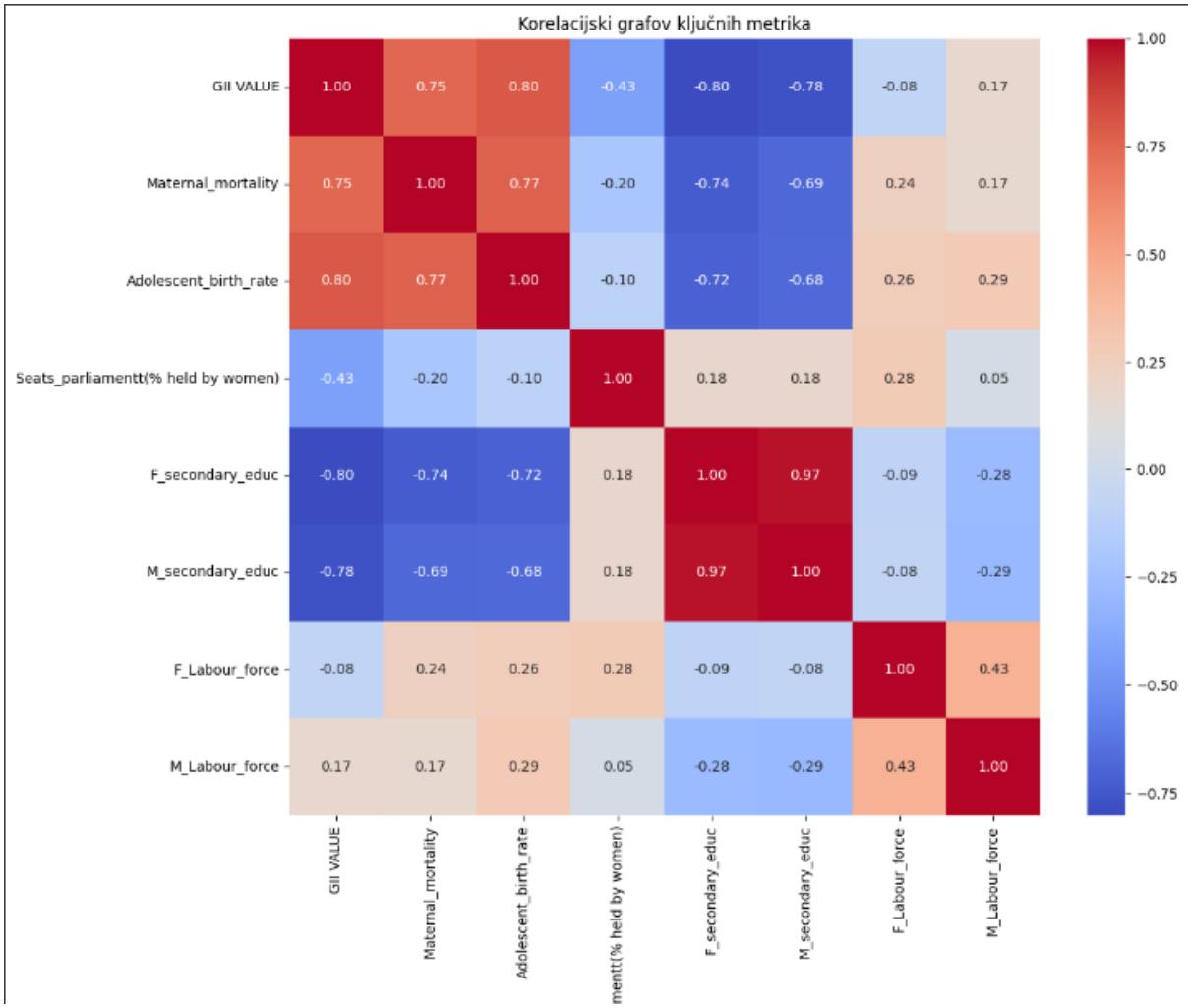


Slika 11. Prikaz Box Plotova ključnih metrika

Analizom grafova zaključuje se slijedeće:

- GII value: graf pokazuje raspon vrijednosti GII-a, neke skupine imaju niže vrijednosti i nekoliko ekstrema s višim vrijednostima
- Maternal_mortality: širok raspon stopa smrtnosti majki, s nekoliko zemalja koje imaju značajno veće stope od ostalih
- Adolescent_birth_rate: značajan raspon u stopama rađanja kod adolescentica
- Seats_parliamentt(% held by women): relativno uski raspon zastupljenosti žena u parlamentima
- F_secondary_educ & M_secondary_educ: širok raspon završenog srednjoškolskog obrazovanja kod oba spola
- F_Labour_force & M_Labour_force: značajan raspon učešća oba spola u radnoj snazi

Korelacijska matrica (slika 12) nije samo alat za identifikaciju odnosa između varijabli, već i ključno sredstvo za razumijevanje kako pojedine dimenzije rodne nejednakosti utječu jedna na drugu.



Slika 12. Korelacijska matrica različitih metrika

- Pozitivne korelacije:

Neke jako pozitivne korelacije su očite i označene su tamnim crvenim kvadratićima. Na primjer, GII VALUE, Maternal_mortality i Adolescent_birth_rate ukazuju na visoke pozitivne korelacije što sugerira da zemlje s visokim vrijednostima GII također imaju tendenciju za visokom stopom smrtnosti majki i visokom stopom nataliteta u adolescenata.

- Negativne korelacije:

Postoje značajne negativne korelacije koje su predstavljene tamno plavim kvadratima. F_secondary_educ i M_secondary_educ pokazuju jake negativne korelacije s GII VALUE, Maternal_mortality i Adolescent_birth_rate što sugerira da zemlje s višom razinom srednjoškolskog obrazovanja obično imaju niže vrijednosti GII-a, stope smrtnosti majki i stope nataliteta u adolescentica.

- Slabije korelacije:

Postoji i nekoliko slabijih korelacija, naznačene svjetlijim nijansama crvene i plave boje. Na primjer, korelacija između Seats_parliamentt(% held by women) i GII VALUE je relativno slaba, što ukazuje na manje izražen odnos između žena u parlamentu i vrijednosti GII.

Kombinirajući sumarne statistike i vizualizacije, deskriptivna analiza pruža uvid u kompleksnost rodne nejednakosti. Identificiranje ključnih obrazaca i korelacije pomaže u razumijevanju kako različiti faktori, poput obrazovanja, zdravstvene njege i političke zastupljenosti, doprinose rodnoj nejednakosti u različitim dijelovima svijeta. Razumijevanje ovih odnosa ključno je za razvoj politika i inicijativa koje mogu pomoći u smanjenju rodne nejednakosti na globalnoj razini.

5. Regresijska analiza

Regresijska analiza predstavlja ključnu metodu za istraživanje između socio-ekonomskih faktora i rodne nejednakosti. Ova analiza omogućava identifikaciju ključnih faktora koji značajno doprinose varijabilnosti u vrijednostima indeksa rodne nejednakosti, tj. GII, među različitim zemljama.

Metoda najmanjih kvadrata (engl. *Ordinary Least Squares*, OLS) koristi se za kvantificiranje odnosa između nezavisnih varijabli (prediktora) i zavisne varijable (GII VALUE) minimizirajući razlike između stvarnih i predviđenih vrijednosti. Ovaj model omogućuje procjenu utjecaja faktora kao što su HDI rangiranje, stopa rađanja kod adolescentica, zastupljenost žena u parlamentu te učešća žena i muškaraca u radnoj snazi na GII. Rezultati modela pružaju važne uvide u faktore koji značajno doprinose povećanju ili smanjenju rodne nejednakosti.

```
X = data[['HDI rank', 'Adolescent_birth_rate', 'Seats_parliamentt(%
               held by women)', 'F_secondary_educ', 'M_secondary_educ',
               'F_Labour_force', 'M_Labour_force']]
y = data['GII VALUE']
X = sm.add_constant(X)
```

Za implementaciju OLS regresijskog modela, odabrane su nezavisne varijable (X) i zavisna varijabla (y), te je dodana konstanta kako bi se procijenio regresijski presjek.

```
model = sm.OLS(y, X).fit()
model.summary()
```

OLS regresijski model se prilagođava podacima i ispisuje sažetak modela, koji uključuje ključne statistike i koeficijente. Te su dobiveni slijedeći sažeti rezultati:

Tablica 2. Sažetak OLS modela

	Vrijednost
Dependent Variable	GII VALUE
R-squared	0.919
Adjusted R-squared	0.916
F-statistic	258.1
Prob (F-statistic)	2.17e-83
Log-Likelihood	246.36
No. Observations	167
AIC	-476.7
Df Residuals	159
BIC	-451.8
Df Model	7

Rezultati OLS regresije pokazuju da model vrlo dobro objašnjava varijacije u GII VALUE, s R-squared vrijednošću od 0.919, što znači da model objašnjava 91.9% varijacije. F-statistika od 258.1, uz izrazito nisku p-vrijednost (2.17e-83), potvrđuje statističku značajnost modela. Ove metrike sugeriraju da su odabrani socio-ekonomski faktori značajni prediktori rodne nejednakosti među zemljama.

Tablica 3. Tablica koeficijenata

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2448	0.058	4.199	0.000	0.130	0.360
HDI rank	0.0019	0.000	9.565	0.000	0.002	0.002
Adolescent_birth_rate	0.0015	0.000	6.840	0.000	0.001	0.002
Seats_parliamentt(% held by women)	-0.0032	0.000	-8.348	0.000	-0.004	-0.002
F_secondary_educ	-0.0009	0.001	-1.258	0.210	-0.002	0.000
M_secondary_educ	0.0003	0.001	0.349	0.728	-0.001	0.002
F_Labour_force	-0.0018	0.000	-5.246	0.000	-0.003	0.001
M_Labour_force	0.0009	0.001	1.480	0.141	0.000	0.002

Ključni zaključci iz OLS regresije:

- HDI rank (koeficijent: 0.0019, $p < 0.000$) i Adolescent_birth_rate (koeficijent: 0.0015, $p < 0.000$) imaju pozitivan i statistički značajan utjecaj na GII VALUE, te ukazuje da više rangiranje po HDI i veća stopa rađanja kod adolescentica povećavaju nejednakost
- Seats_parliamentt (% held by women) (koeficijent: -0.0032, $p < 0.000$) i F_Labour_force (koeficijent: -0.0018, $p < 0.000$) imaju negativan i statistički značajan utjecaj na GII VALUE, što sugerira da veća zastupljenost žena u parlamentu i radnoj snazi smanjuje rodnu nejednakost
- F_secondary_educ (koeficijent: -0.0009, $p = 0.210$), M_secondary_educ(koeficijent: 0.0003, $p = 0.728$) i M_Labour_force(koeficijent: 0.0009, $p = 0.141$) nisu statistički značajni na nivou od 5%, što znači da njihov utjecaj na GII VALUE nije dovoljno značajan u ovom modelu

Dijagnostički tekstovi koriste se za procjenu valjanosti i stabilnosti regresijskog modela kroz ključne metrike. Za provjeru normalnosti reziduala koriste se Omnibus i Jarque-Bera testovi, a poželjne p-vrijednosti su iznad 0.05 kako bi se potvrdila ta normalnost, što je važno za ispravne procjene modela. Durbin-Watson test s idealnim rasponom između 1.5 i 2.5 ukazuje na odsutnost autokorelacije među rezidualima. Skewness blizu 0 i Kurtosis oko 3 sugeriraju simetričnu i normalnu distribuciju, dok Condition Number ispod 30 pokazuje nisku multikolinearnost, osiguravajući stabilne i pouzdane rezultate modela [10].

Tablica 4. Dijagnostički testovi

Test	Vrijednost	Vjerojatnost
Omnibus	22.380	0.000
Durbin-Watson	2.099	-
Jarque-Bera (JB)	43.062	4.46e-10
Skew	0.630	-
Kurtosis	5.145	-
Condition Number	2.22e+03	-

- Omnibus Test i Jarque-Bera Test

P-vrijednosti (0.000 i 4.46e-10) pokazuju statistički značajna odstupanja od normalne distribucije reziduala, što sugerira da reziduali nisu normalno raspoređeni i može ukazivati na prisutnost outliera ili neadekvatnost modela u objašnjavanju varijabilnosti podataka.

- Durbin-Watson Test

Vrijednost od 2.099 ukazuje na odsutnost značajne autokorelaciješto je pozitivan znak za stabilnost modela.

- Skewness (asimetrija) i Kurtosis (špicastost)

Skewness od 0.630 ukazuje na umjerenu pozitivnu asimetriju, dok Kurtosis od 5.145 sugerira povećanu šiljatost distribucije u odnosu na normalnu.

- Condition Number

Condition Number od 2.22e+03 sugerira prisutnost značajne multikolinearnosti među prediktorima, što može destabilizirati model i utjecati na pouzdanost koeficijenata.

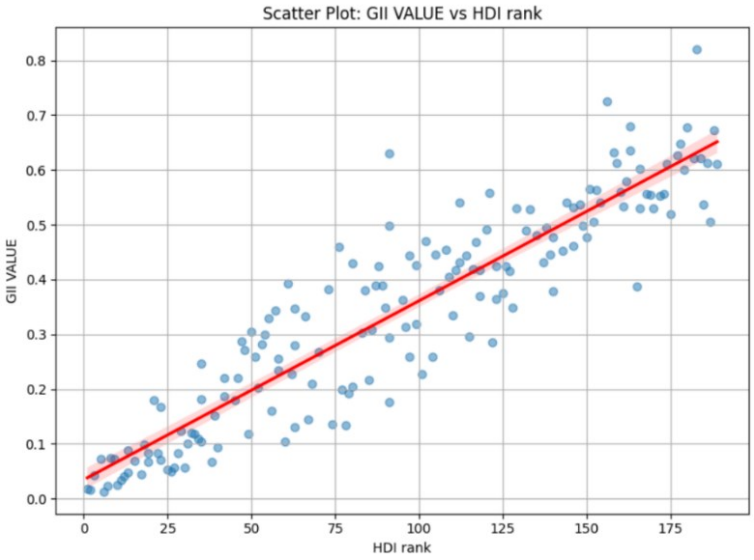
5.1. Vizualizacija regresijske analize

Za bolje razumijevanje odnosa između nezavisnih varijabli i GII VALUE kreirani su dijagrami raspršenosti (engl. *scatter plots*) s regresijskim linijama. Oni prikazuju odnos između nezavisnih varijabli i GII vrijednosti, s uključenim trend linijama koje dodatno vizualiziraju smjer i snagu odnosa.

```
independent_variables = [
    "HDI rank",
    "Adolescent_birth_rate",
    "Seats_parliamentt(% held by women)",
    "F_secondary_educ",
    "M_secondary_educ",
    "F_Labour_force",
    "M_Labour_force"
]
for variable in independent_variables:
    plt.figure(figsize=(8, 6))
    sns.regplot(
        x=variable,
        y='GII VALUE',
```

```
    data=data,
    line_kws={"color": "red"},
    scatter_kws={"alpha": 0.5}
)
```

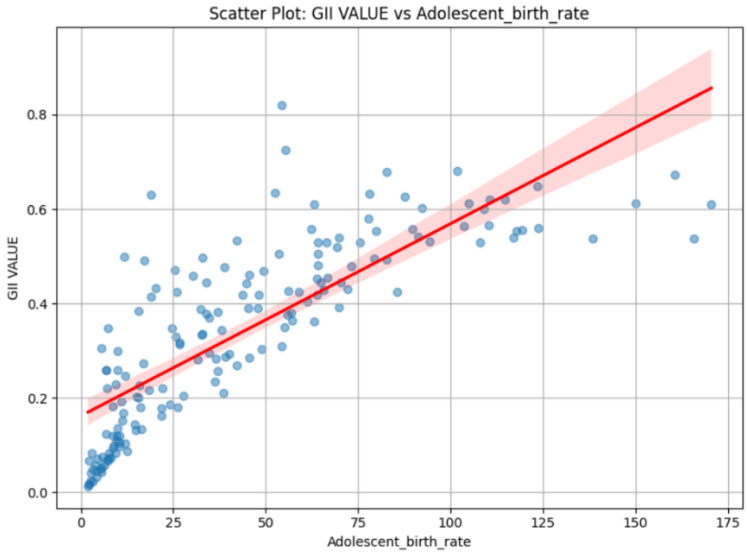
- HDI rang:



Slika 13. Dijagram raspršenosti: GII VALUE i HDI rank

Zemlje s nižim vrijednostima indeksa ljudskog razvoja (viši rangom HDI-a) obično imaju veću rodnu nejednakost (slika 13). Poboljšanje ukupnog ljudskog razvoja može doprinijeti smanjenju rodne nejednakosti.

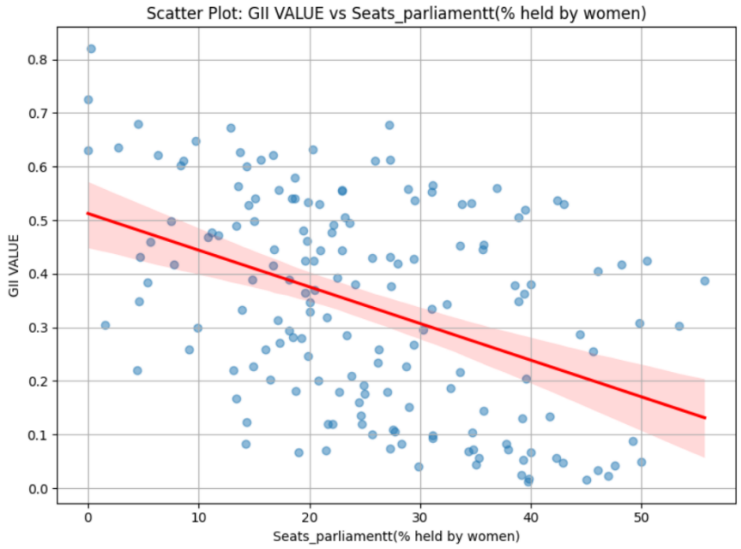
- Stopa nataliteta adolescentica:



Slika 14. Dijagram raspršenost: GII VALUE i Adolescent_birth_rate

Postoji jaka pozitivna korelacija između GII VALUE i Adolescent_birth_rate, što znači da su veće stope nataliteta povezane s većom rodnom nejednakošću (slika 14). Smanjenje stope nataliteta adolescenata moglo bi značajno poboljšati rodnu ravnopravnost.

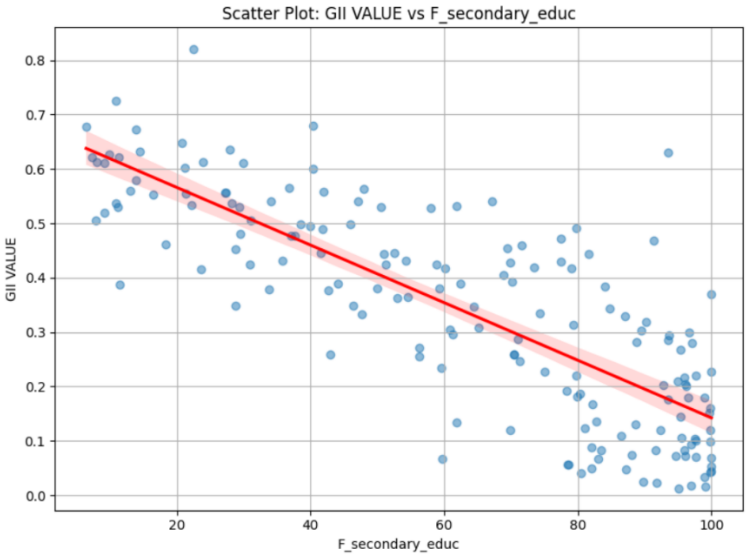
- Zastupljenost žena u parlamentu:



Slika 15. Dijagram raspršenost: GII VALUE i Seats_parliamentt(% held by women)

Postoji negativna korelacija između GII VALUE i Seats_parliamentt(% held by women), što znači da je veća politička zastupljenost žena u parlamentu povezan s manjom rodnom nejednakošću (slika 15). Poticanje većeg broja žena na sudjelovanje u politici može pomoći poboljšanju rodne ravnopravnosti.

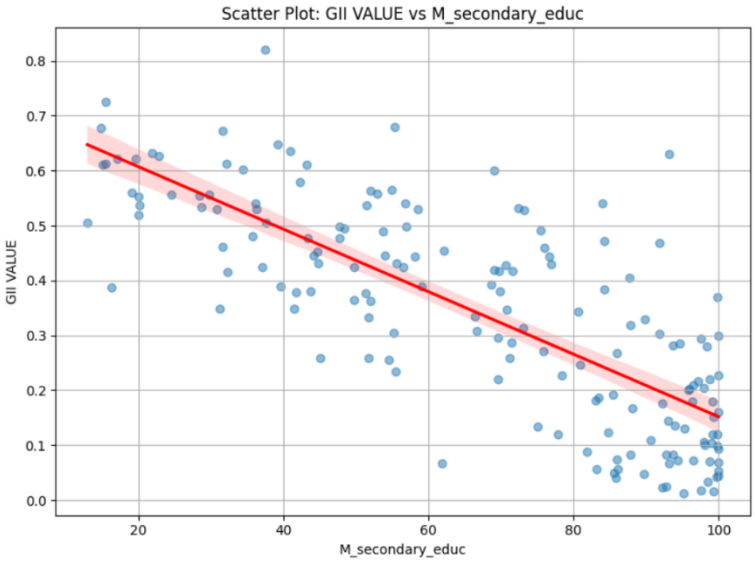
- Žensko srednjoškolsko obrazovanje:



Slika 16. Dijagram raspršenost: GII VALUE i F_secondary_educ

Između GII VALUE i F_secondary_educ postoji jako negativna korelacija, tj. žensko srednjoškolsko obrazovanje snažno je povezano s manjom rodnom nejednakošću, što ukazuje da je promicanje ženskog obrazovanja ključno za smanjenje rodne diskriminacije (slika 16).

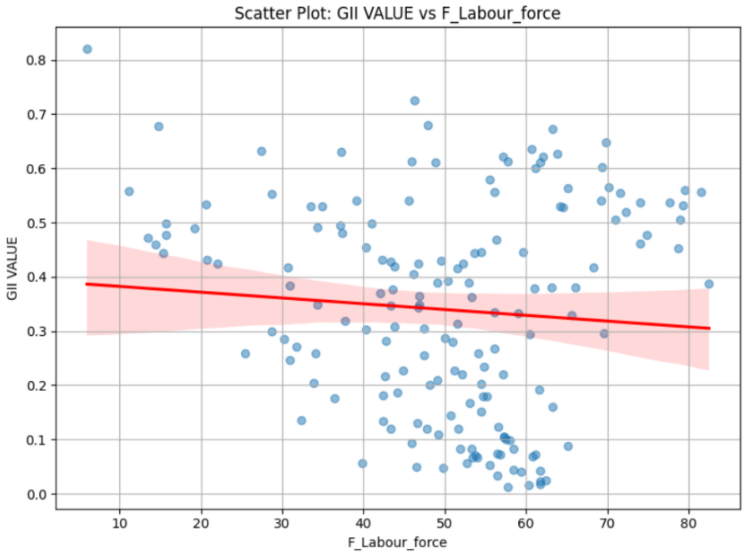
- Muško srednjoškolsko obrazovanje:



Slika 17. Dijagram raspršenost: GII VALUE i M_secondary_educ

Ova varijabla nije značajna, što sugerira da srednjoškolsko obrazovanje muškaraca nema značajan utjecaj na rodnu nejednakost u ovome modelu (slika 17).

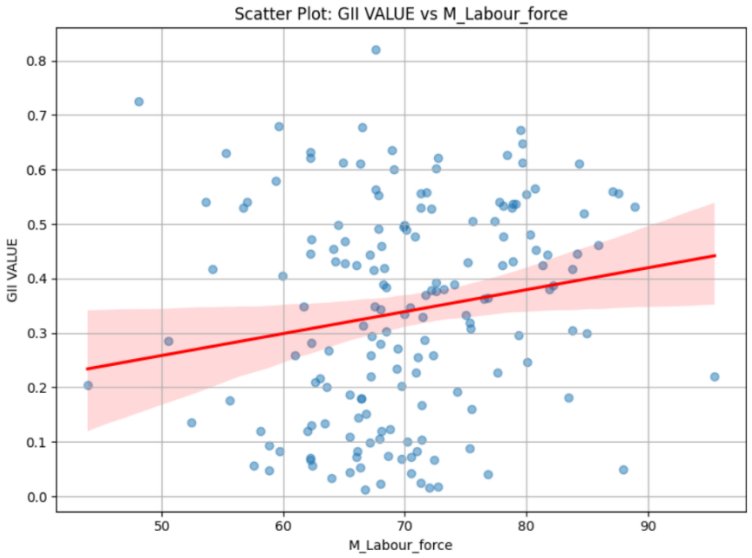
- Sudjelovanje žena u radnoj snazi:



Slika 18. Dijagram raspršenost: GII VALUE i F_Labour_force

Blago negativna korelacija postoji između GII VALUE i F_Labour_force, što znači da je veća participacija žena u radnoj snazi povezana s manjom rodnom nejednakošću (slika 18). Poticanje žena na sudjelovanje u radnoj snazi moglo bi poboljšati ravnopravnost spolova.

- Sudjelovanje muškaraca u radnoj snazi:



Slika 19. Dijagram raspršenost: GII VALUE i M_Labour_force

Iako ova varijabla pokazuje pozitivnu povezanost, ona nije statistički značajna, što ukazuje na to da učešće muškaraca u radnoj snazi nema značajan utjecaj na GII VALUE u ovom modelu (slika 19).

Rezultati regresijske analize ukazuju na ključne faktore koji utječu na rodnu nejednakost. Identifikacija ovih faktora može poslužiti kao temelj za donošenje politika usmjerenih na smanjenje rodne diskriminacije, s posebnim naglaskom na unaprjeđenje obrazovanja, participaciju žena u političkom radu te smanjenje stope nataliteta kod adolescentica.

6. Klasteriranje i detekcija anomalija

Klasteriranje i detekcija anomalija omogućuju grupiranje zemalja prema sličnim karakteristikama rodne nejednakosti te identifikaciju zemalja koje se značajno razlikuju od ostalih [8]. Kroz klaster analizu moguće je identificirati skupine zemalja sa sličnim profilima rodne nejednakosti, dok detekcija anomalija omogućava prepoznavanje onih zemalja koje značajno odstupaju od uobičajenih obrazaca.

6.1. Klasteriranje

Klaster analiza, tj. klasteriranje, je nenadgledana metoda učenja koja se koristi za grupiranje podataka u klastere na temelju sličnosti među njima. Ove skupine, ili klasteri, formiraju se na temelju unutarnjih struktura podataka, bez unaprijed definiranih klasa. Klasteriranje može otkriti prirodne grupe unutar podataka koje reflektiraju određene mehanizme djelovanja u domeni iz koje podaci potječu, što rezultira time da su neki podaci sličniji jedni drugima nego preostalim podacima [8]. Korištenjem ove tehnike cilj je identificirati grupe zemalja sa sličnim profilima rodne nejednakosti, što može pomoći u prepoznavanju obrazaca koje ukazuju na potencijalnu diskriminaciju.

```
label_encoder = LabelEncoder()
data['HUMAN DEVELOPMENT'] = label_encoder.fit_transform(data['HUMAN DEVELOPMENT'])
features = data.columns.drop(['HDI rank', 'Country'])
```

Prije nego što se provede klasteriranje, potrebno je obraditi podatke kako bi bili u ispravnom formatu. To uključuje kodiranje kategoričkih varijabli i skaliranje značajki. Stupac HUMAN DEVELOPMENT transformiran je u numeričke vrijednosti koristeći *LabelEncoder*. Značajke koje se trebaju skalirati odabrane su iz skupa podataka, s izuzetkom stupaca HDI rank i Country, koje se ne skaliraju zbog već odgovarajućeg formata.

```
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data[features])
scaled_df = pd.DataFrame(scaled_data, columns=features)
```

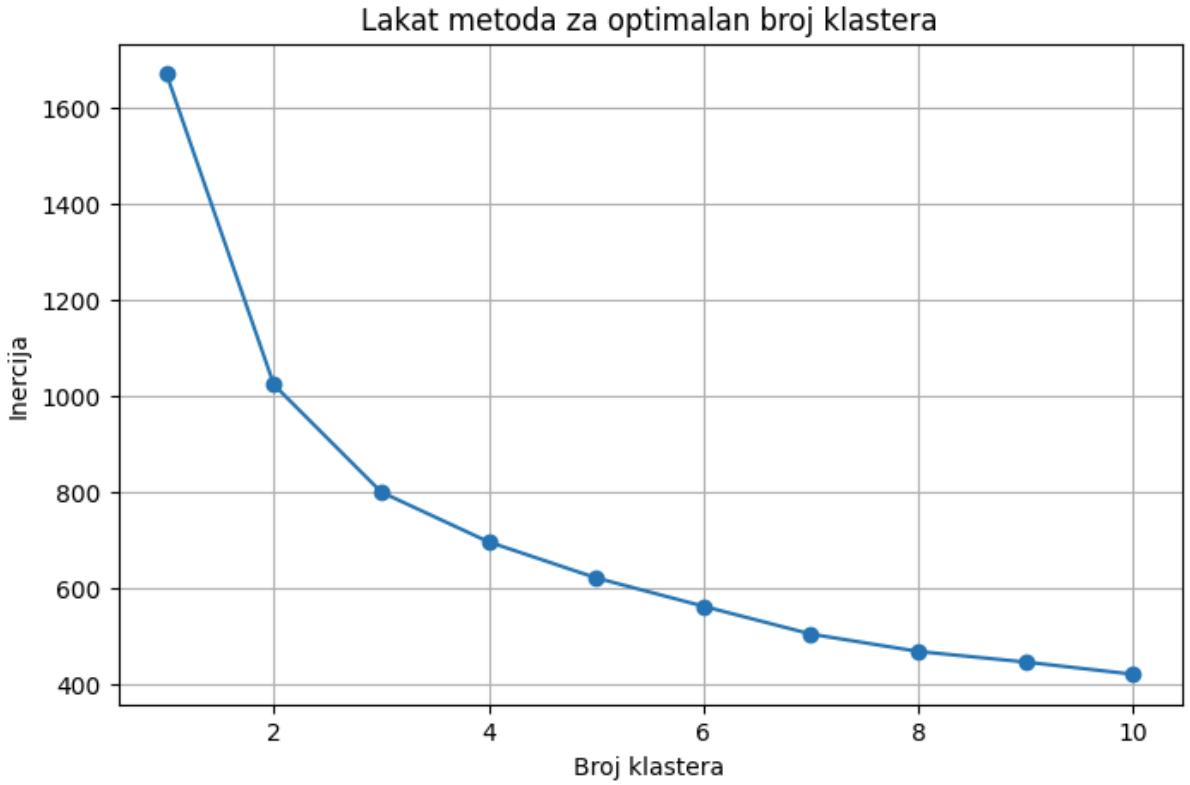
Da bi se osiguralo da svaka značajka jednako doprinosi izračunu udaljenosti u klaster algoritmu, značajke su standardizirane koristeći *StandardScaler* kako bi imale srednju vrijednost 0 i standardnu devijaciju 1. Zatim se skalirani podaci organiziraju u novom DataFrame-u za lakšu manipulaciju i analizu.

Za klasterizaciju je odabran K-Means algoritam zbog svoje efikasnosti i jednostavnosti. K-Means algoritam dijeli podatke u K različitih klastera tako da minimizira unutar-klastersku varijancu (inerciju). Ovaj algoritam radi tako što najprije nasumično odabere K točaka kao središta klastera. Zatim se svi podaci dodjeljuju svom najbližem klaster centru prema standardnoj euklidskoj udaljenosti. Sljedeći korak je izračunavanje centroida, odnosno srednje vrijednosti, za svaki klaster, što postaje novo središte klastera. Proces se ponavlja sve dok se podaci više ne premještaju između klastera, što znači da su klaster centri stabilizirani.

```
inertia = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_df)
    inertia.append(kmeans.inertia_)
```

Za određivanje optimalnog broja klastera korištena je lakat metoda (engl. *Elbow Method*). Ova metoda pomaže identificirati optimalan broj klastera iscrtavanjem zbroja kvadrata udaljenosti unutar klastera za različite brojeve klastera. „Lakat“ je točka na gdje se inercija počinje naglo smanjivati, označavajući optimalan broj klastera. Ova točka predstavlja balans između složenosti modela i njegove učinkovitosti u grupiranju podataka.

Dijagram lakat metode prikazuje primjetno savijanje kod 3 klastera, što označuje da je 3 razuman izbor za broj klastera (slika 20).



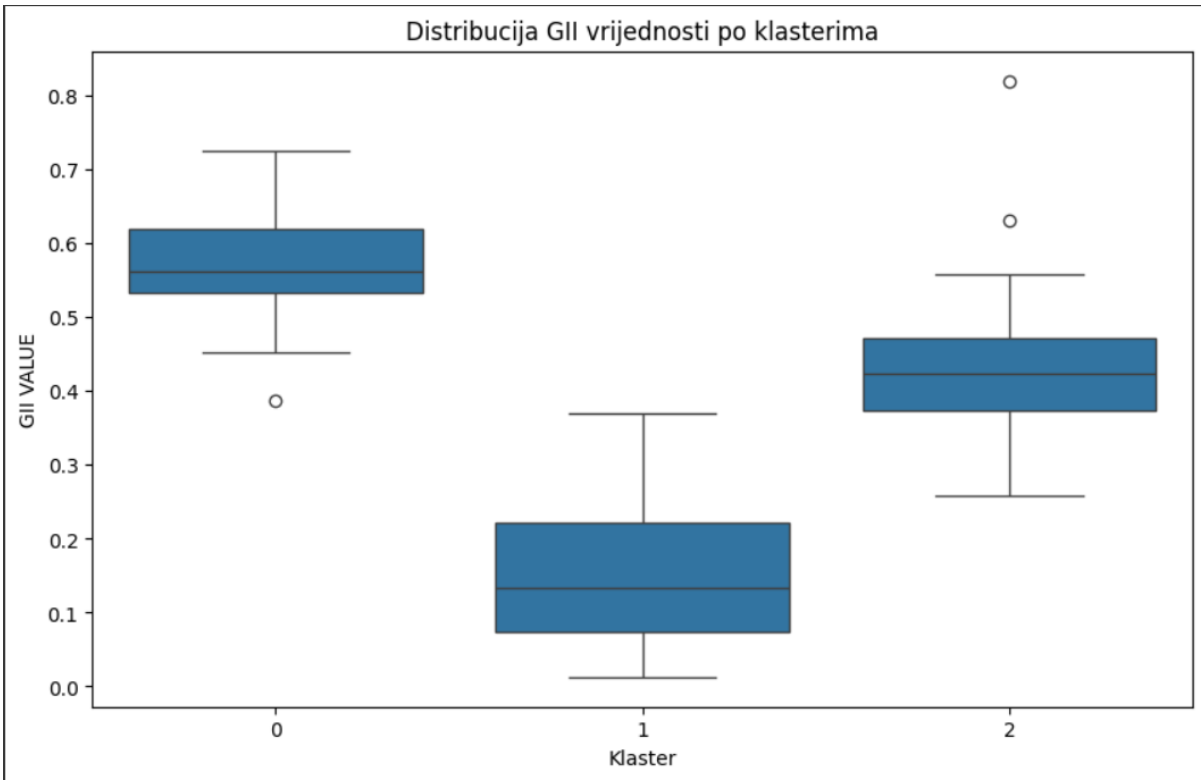
Slika 20. Dijagram lakat metode

```
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(scaled_df)
data['Cluster'] = clusters
```

Nakon što je određen optimalni broj klastera pomoću metode lakta, sljedeći korak je primijeniti K-Means algoritam s tim brojem klastera na skalirane podatke. U ovom slučaju, optimalni broj klastera je tri, što znači da će algoritam grupirati zemlje u tri različite skupine. Korištenje

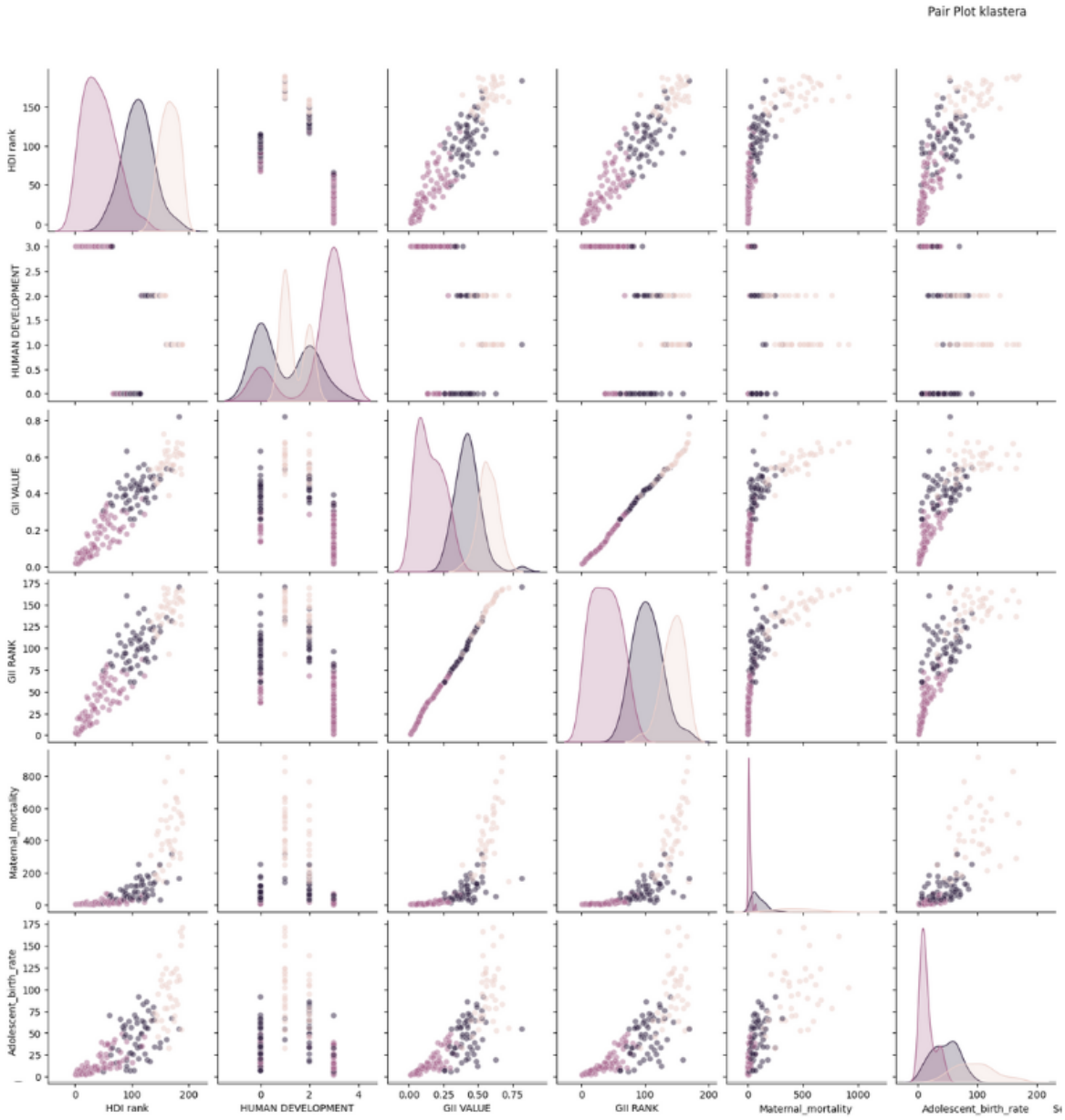
funkcije *fit_predict* na skaliranim podacima omogućava modelu da nauči koji podatci pripadaju kojem klasteru, te na temelju toga svakoj zemlji dodijeli oznaku klastera (0, 1 ili 2). Ove oznake se zatim pohranjuju u novi stupac pod nazivom „Cluster“ u originalnom podatkovnom okviru, što nam pomaže razumjeti obrasce i grupiranja među zemljama.

Nakon primjene K-Means algoritma, provodi se analiza distribucije GII vrijednosti po klasterima kako bi se utvrdilo postoje li značajne razlike u rodnoj nejednakosti među različitim skupinama zemalja.



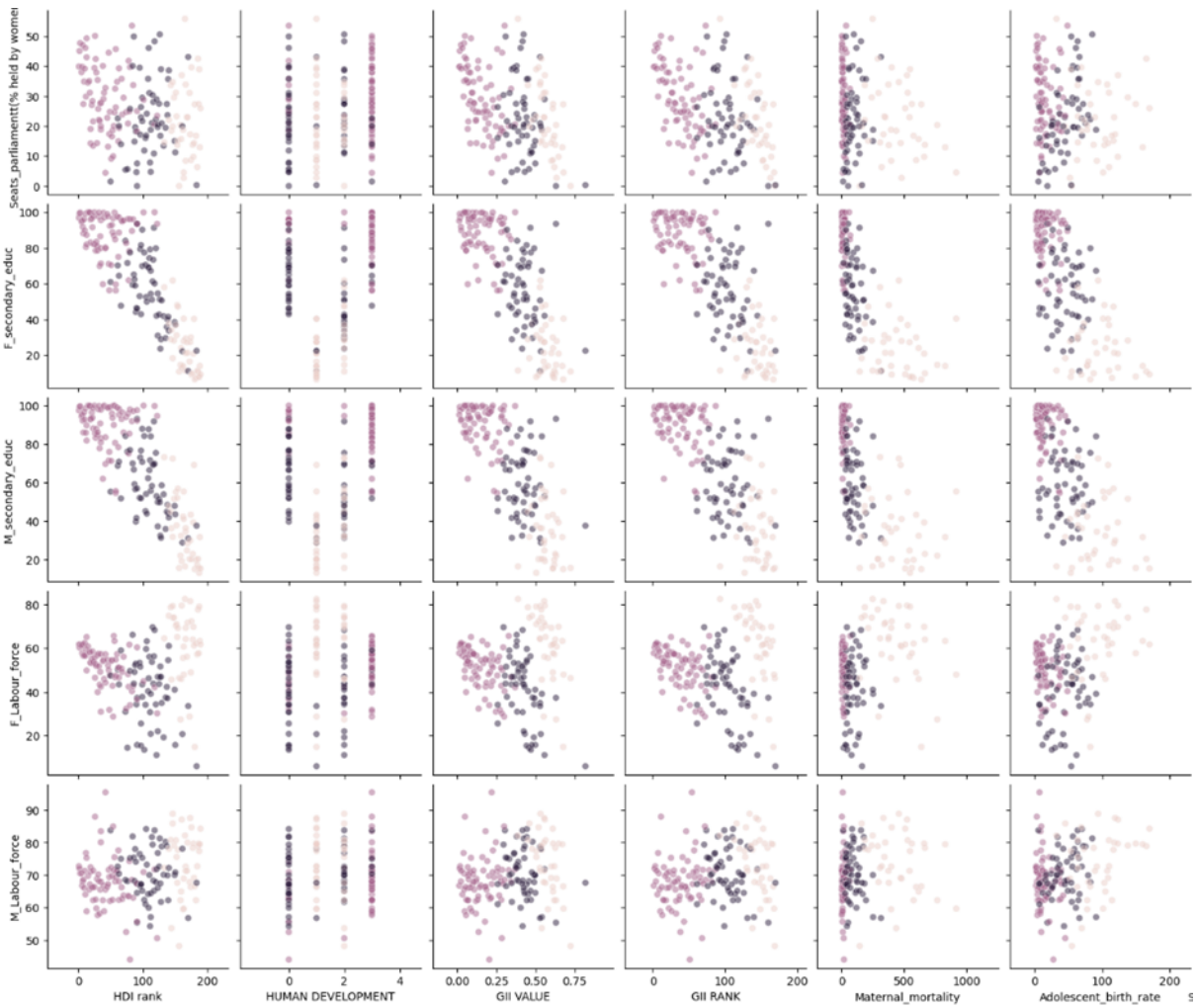
Slika 21. Box plotovi distribucije GII vrijednosti po klasterima

Na slici 21 je prikazana raspodjela GII vrijednosti unutar svakog klastera. Klasteri s visokim GII vrijednostima ukazuju na zemlje s većom rodnom nejednakošću, dok klasteri s niskim GII vrijednostima pokazuju zemlje koje imaju nižu razinu rodne nejednakosti.

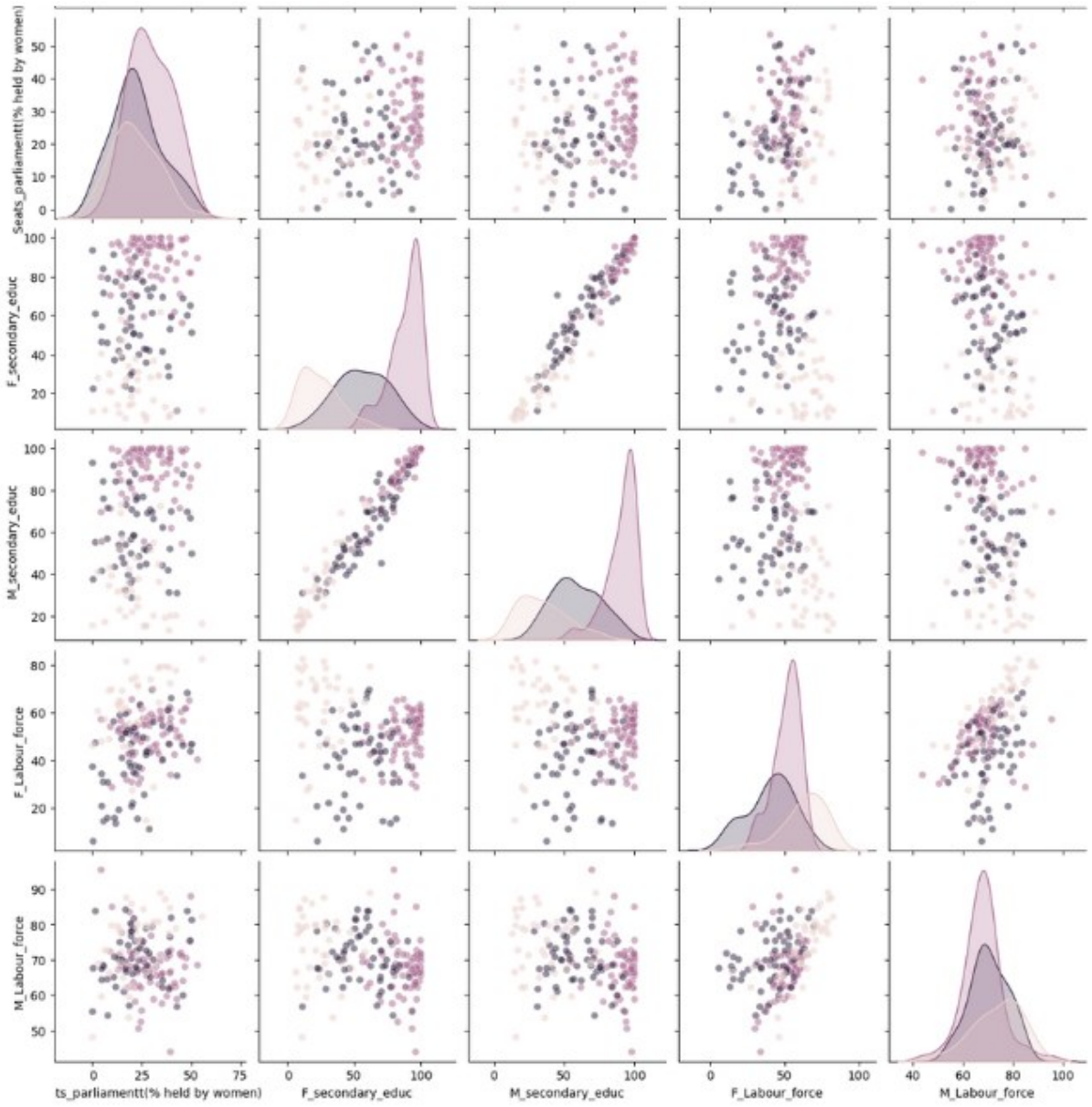


Slika 22a. Pair Plot klastera

Slika 22b. Pair Plot klastera



Slika 22c. Pair Plot klastera



Slika 22d. Pair Plot klastera

Na slikama 22a, 22b, 22c i 22d prikazana je parcijalna vizualizacija odnosa između različitih varijabli u skupu podataka dobivenih K-Means algoritmom, gdje su zemlje grupirane u tri klastera označena različitim bojama (0, 1 i 2). Svaki točkasti grafikon prikazuje odnos između dvije varijable, dok dijagonalni grafikon prikazuje distribuciju svake varijable unutar klastera.

- Jasna separacija klastera: u nekoliko kombinacija varijabli vidljiva je jasna separacija između klastera, što sugerira da su varijable ključne u definiranju sličnosti među zemljama:
  - Maternal_mortality i Adolescent_birth_rate su dvije varijable koje pokazuju značajnu razliku između klastera. Klaster 0 se ističe s najvišim stopama maternalne smrtnosti i

adolescentnih trudnoća, što ukazuje na zemlje s većim zdravstvenim rizicima. Klaster 2 obuhvaća zemlje s najnižim stopama, dok je klaster 1 u sredini.

- HUMAN_DEVELOPMENT pokazuje specifične obrasce gdje zemlje u klasteru 0 imaju niži stupanj razvoja, dok zemlje u klasteru 2 pokazuju visoku razinu ljudskog razvoja. Klaster 1 predstavlja zemlje srednjeg razvoja.

- Preklapanje klastera: kod nekih varijabli primjerno je preklapanje klastera, što sugerira da su te varijable manje učinkovite u razdvajanju zemalja.
  - Seats_parliamentt(% held by women) ne prikazuje jasnu separaciju između klastera, što znači da zastupljenost žena u parlamentu nije presudna za razlikovanje grupa zemalja u ovom skupu podataka.

  - slično, F_Labour_force, pokazuje određeno preklapanje među klasterima, što upućuje na to da i stopa sudjelovanja žena u radnoj snazi nije ključna za razdvajanje žena prema klasterima.

Vizualizacija parcijalnih dijagrama jasno ističe varijable koje snažno utječu na formiranje klastera, kao što su maternalna smrtnost, adolescentne trudnoće i ljudski razvoj. Istovremeno, neke varijable poput zastupljenosti žena u parlamentu i sudjelovanja žena u radnoj snazi pokazuju veće preklapanje među klasterima, sugerirajući da su manje učinkovite u razdvajanju zemalja na temelju analiziranih karakteristika. Ova analiza pomaže u razumijevanju koji su faktori ključni za grupiranje zemalja u ovom kontekstu i kako se oni međusobno odnose.

6.2. Detekcija anomalija

Za otkrivanje anomalija u podacima, koristi se algoritam izolacijskih šuma (engl. *Isolation Forest*) kojim se nastoji izolirati stršeće vrijednosti iz skupa podataka. Ovaj algoritam posebno je učinkovit za identifikaciju odstupanja u visokodimenzionalnim skupovima podataka. Algoritam izolacijskih šuma radi na principu izoliranih točaka u skupu podataka nasumičnim odabirom značajki i vrijednosti za podjelu pri čemu su anomalije obično izolirane kraćim stazama jer su lakše odvojive.

Ključni koraci:

- Nasumično poduzorkovanje
  - ❖ Odabir podskupa točaka i značajki.
- Rekurzivno particioniranje
  - ❖ Za svaku odabranu značajku, nasumično se bira vrijednost za podjelu.
  - ❖ Podaci se dijele na lijeve i desne grane na temelju vrijednosti podjele.
  - ❖ Postupak se ponavlja rekurzivno kako bi se stvorila izolacijska stabla.
- Izračun duljine puta

- ❖ Bilježi se broj podjela potrebnih za izoliranje svake točke (duljina puta).
- ❖ Anomalije imaju kraće putove jer se lakše izoliraju.

- Ocjena anomalije
  - ❖ Ocjena anomalije izračunava se na temelju prosječne duljine puta u izolacijskim stablima.
  - ❖ Točke s kraćim prosječnim putevima dobivaju veću ocjenu, što ukazuje na veću vjerojatnost da su anomalije.

```
iso_forest = IsolationForest(contamination=0.05, random_state=42)
anomalies = iso_forest.fit_predict(scaled_df)
data['Anomaly'] = anomalies
anomalies_detected = data[data['Anomaly'] == -1]
anomalies_detected
```

Primjenjuje se algoritam izolacijske šume na cijeli skup podataka kako bi se otkrile globalne anomalije.

Model *IsolationForest* je inicijaliziran s parametrima *contamination=0.05* koji označava očekivanih oko 5% anomalija u skupu podataka te *random_state=42* za reproducibilnost rezultata. Nakon primjene modela podaci označeni s „-1“ smatraju se anomalijama, dok vrijednost ’1’ označava normalan podatak. Rezultati detekcije anomalija dodaju se u originalni DataFrame kao novi stupac Anomaly te se prikazuju (slika 23).

	HDI rank	Country	HUMAN DEVELOPMENT	GII VALUE	GII RANK	Maternal_mortality	Adolescent_birth_rate	Seats_parliamentt(% held by women)	F_secondary_educ	M_secondary_educ	F_Labour_force	M_Labour_force	Cluster	Anomaly
92	91	Tonga	0	0.631	160.0	52.0	19.0	0.0	93.5	93.1	37.3	55.3	2	-1
156	156	Papua New Guinea	2	0.725	169.0	145.0	55.3	0.0	10.8	15.5	46.3	48.1	0	-1
157	158	Mauritania	2	0.632	161.0	766.0	78.0	20.3	14.5	21.9	27.4	62.2	0	-1
163	163	Nigeria	1	0.680	168.0	917.0	101.7	4.5	40.4	55.3	47.9	59.6	0	-1
164	165	Rwanda	1	0.388	93.0	248.0	32.4	55.7	11.4	16.3	82.5	82.2	0	-1
179	180	Afghanistan	1	0.678	167.0	638.0	82.6	27.2	6.4	14.9	14.8	66.5	0	-1
182	183	Yemen	1	0.820	170.0	164.0	54.4	0.3	22.4	37.5	6.0	67.6	2	-1
184	185	Mozambique	1	0.537	136.0	289.0	165.8	42.4	10.8	20.2	77.7	78.9	0	-1
187	188	Central African Republic	1	0.672	166.0	829.0	160.5	12.9	13.9	31.6	63.3	79.5	0	-1

Slika 23. Rezultati detekcije anomalija

Identifikacija anomalija pomoću izolacijskih šuma omogućila je prepoznavanje zemalja koje predstavljaju ekstremne slučajeve rodne nejednakosti. Ove zemlje, kao što su Tonga, Papua Nova Gvineja, Mauritanija, Nigerija, Ruanda, Afganistan, Jemen, Mozambik i Centralnoafrička Republika zahtijevaju poseban fokus i prilagođene intervencije kako bi se smanjila diskriminacija i unaprijedila rodna ravnopravnost.

Istaknute zemlje pokazuju značajne devijacije u karakteristikama kao što su smrtnost povezane s trudnoćom, stope adolescentnih trudnoća ili pokazatelja rodne nejednakosti u usporedbi s drugim zemljama.

6.3. Vizualizacija klasteriranja i detekcije anomalija

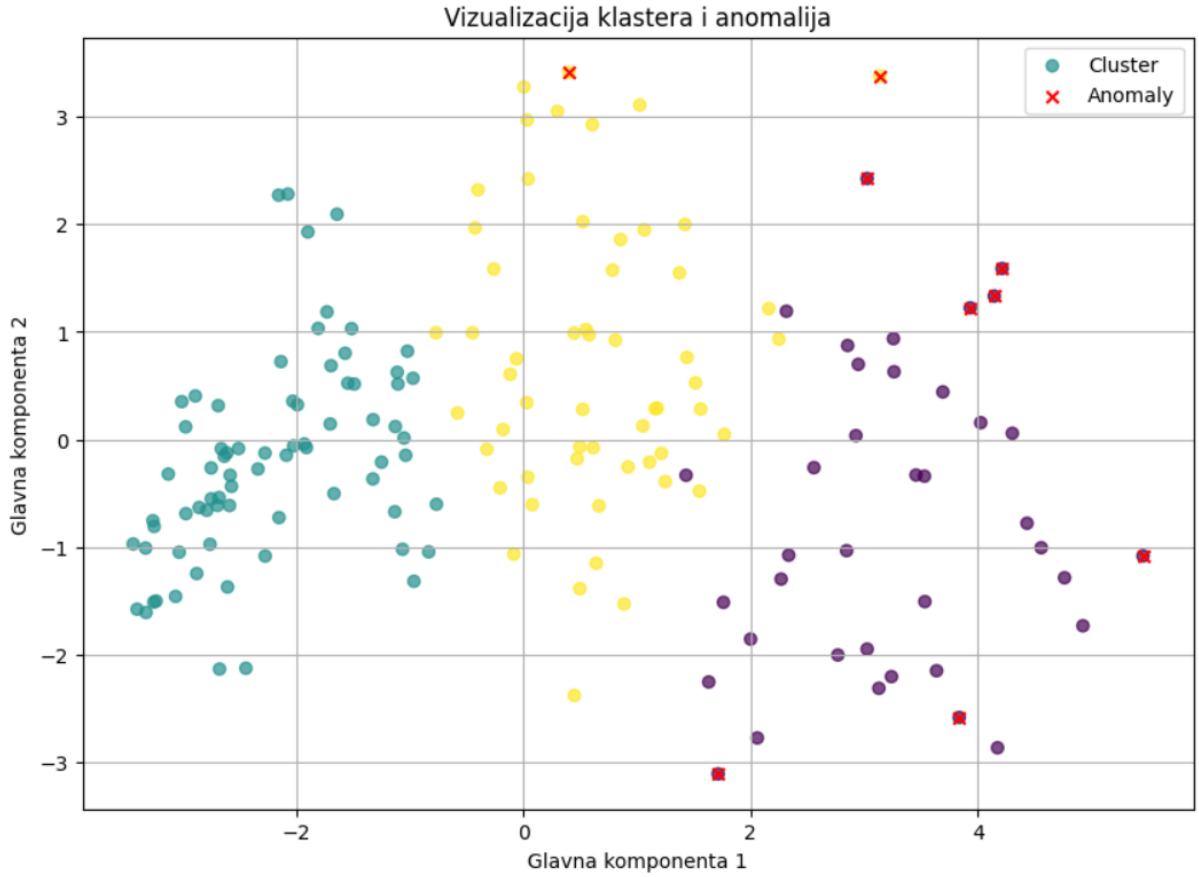
Za vizualizaciju klastera i anomalija, koristi se PCA kako bi smanjili dimenzije podataka na dvije glavne komponente. Na grafu će točke biti obojane s njihovim odgovarajućim bojama klastera te označeni ako su identificirane kao anomalije.

```
pca = PCA(n_components=2)
pca_result = pca.fit_transform(scaled_df)
```

Inicijalizira se PCA algoritam s dvije komponente kako bi se smanjila visokodimenzijalnost podataka. Zatim se PCA model koristi na skaliranim podacima i transformira podatke u dvije glavne komponente, bilježeći najznačajniju varijaciju u podacima.

```
pca_df = pd.DataFrame(data=pca_result, columns=['PC1', 'PC2'])
pca_df['Cluster'] = clusters
pca_df['Anomaly'] = anomalies
```

Kako bi se organizirali rezultati PCA i uključile oznake klastera i anomalija za svaku podatkovnu točku kreira se novi DataFrame pod nazivom „pca_df“ sa stupcima za dvije komponente „PC1“ i „PC2“.



Slika 24. Vizualizacija klastera i anomalija

Vizualizacija klastera pomoću raspršenih dijagrama nakon PCA transformacije jasno prikazuje razdvajanje zemalja u tri klastera, gdje svaka boja predstavlja jedan klaster. Anomalije su jasno označene crvenim križićima, koji označavaju zemlje koje značajno odstupaju od uobičajenih obrazaca (slika 24). Ova vizualizacija omogućava bolje razumijevanje karakteristika koje grupiraju zemlje u određene klastere, kao i prepoznavanje onih zemalja koje se ističu po ekstremnim vrijednostima.

7. Klasifikacija

Klasifikacija je tehnika koja koristi modele za predviđanje diskretnih klasa na temelju ulaznih podataka. Ovi modeli, poznati kao klasifikatori, omogućuju da se razumije struktura podataka i predviđa kategorije za nove uzorke [11]. Klasifikatori, kao što su algoritmi slučajnih šuma (engl. *Random Forest*), stabla odlučivanja (engl. *Decision Tree*) i naivni Bayes, koriste se za prepoznavanje obrazaca u podacima i predviđanje kategorizacije novih podataka na temelju naučenih modela. Klasifikacija se široko primjenjuje u raznim područjima, uključujući detekciju prijevara, medicinskih dijagnoza, marketinških strategija pa tako može služiti i za otkrivanje diskriminacije kroz analizu rodne nejednakosti.

```
non_numeric_columns = data.select_dtypes(include=['object']).columns
label_encoder = LabelEncoder()
for column in non_numeric_columns:
    data[column] = label_encoder.fit_transform(data[column])
```

Nenumeričke stupci se identificiraju i kodiraju pomoću *LabelEncoder*, čime se tekstualne vrijednosti pretvaraju u numeričke oznake koje model može koristiti.

```
features = data.drop(['GII VALUE', 'Country', 'GII RANK'], axis=1)
target = data['GII VALUE']
scaler = StandardScaler()
features_scaled = scaler.fit_transform(features)
```

Kako je ciljana varijabla GII VALUE, potrebno ju je isključiti iz selekcije zajedno s GII RANK, a Country je nenumerički stupac koji nema izravnog utjecaja na klasifikaciju. Preostale značajke se standardiziraju pomoću *StandardScaler*, koji mijenja vrijednosti značajki tako da imaju srednju vrijednost 0 i standardnu devijaciju 1.

```
bins = [0, 0.3, 0.6, 1]
labels = ['Low', 'Medium', 'High']
data['GII_category'] = pd.cut(data['GII VALUE'], bins=bins, labels=labels)
```

Ciljna varijabla GII VALUE podijeljena je u tri kategorije (Low, Medium, High) što omogućuje klasifikacijski pristup analizi.

Za klasifikaciju se koristi algoritam slučajnih šuma, jedan od najpopularnijih algoritama zbog svoje robusnosti. Algoritam je ansambl metoda koja koristi više stabala za klasifikaciju. Svako stablo je trenirano na različitim podskupovima podataka i koristi nasumične podskupove značajki za podjelu čvorova. Konačna odluka temelji se na većinskom glasanju svih stabala u šumi.

```
X = features_scaled
y = data['GII_category']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
classifier = RandomForestClassifier(n_estimators=100, random_state=42)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
print(classification_report(y_test, y_pred))
accuracy = accuracy_score(y_test, y_pred)
```

Podaci su podijeljeni u skup za učenje i skup za ispitivanje, u omjeru 80-20. Model je evaluiran korištenjem izvještaja o klasifikaciji koji uključuje preciznost, odziv i F1 mjeru za svaku kategoriju. Ispisuje se izvještaj o klasifikaciji i točnost modela (slika 25).

```
              precision    recall  f1-score   support

        High       0.50      1.00      0.67         1
         Low       0.83      0.94      0.88        16
      Medium       0.93      0.76      0.84        17

    accuracy                           0.85        34
   macro avg       0.75      0.90      0.80        34
weighted avg       0.87      0.85      0.85        34

Točnost: 0.8529411764705882
```

Slika 25. Predviđanje i evaluacija modela

Model postiže točnost od 85.29%, što ukazuje na ispravno predviđanje klase GII vrijednosti u 85,29% slučajeva.

7.1. Vizualizacija klasifikacije

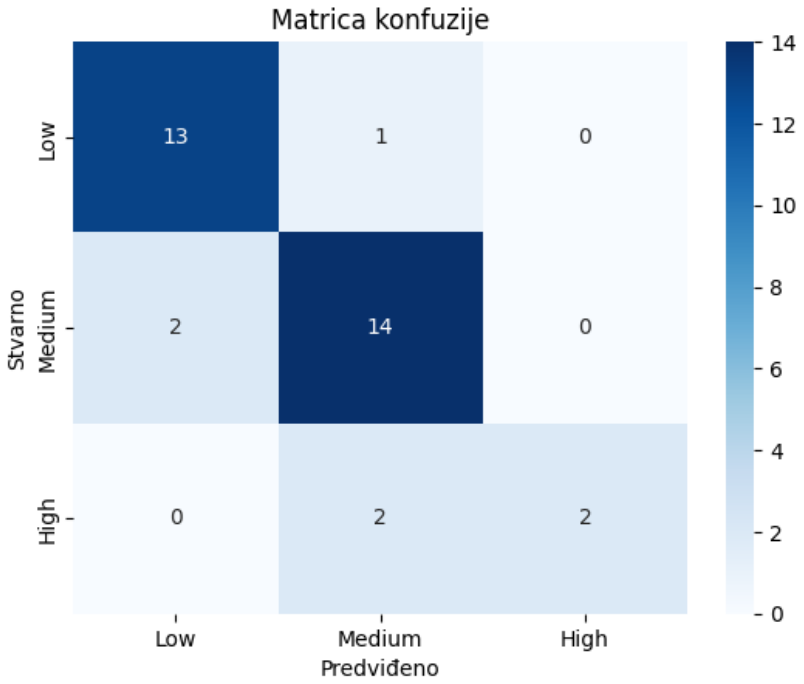
Ciljana varijabla GII_category podijeljena je u tri kategorije: Low, Medium i High. Da bismo razumjeli koliko su klase uravnotežene, koristi se funkcija *value_counts()* u Pandas za pregled distribucije svake klase.

```
GII_category
Medium    77
Low       72
High      18
Name: count, dtype: int64
```

Slika 26. Rezultat distribucije klasa

Ova raspodjela pokazuje da su klase Medium i Low relativno ravnotežno zastupljene, dok je klasa High značajno manje zastupljena u skupu podataka (slika 26).

Kreira se matrica konfuzije za vizualizaciju učinka klasifikatora. Na njoj se prikazuje broj točnih i netočnih predviđanja za svaku klasu, a radi bolje inerpretabilnosti koristi se toplinska karta (slika 27).



Slika 27. Matrica konfuzije

Matrica konfuzije i toplinska karta pružaju detaljan prikaz učinka klasifikatora. Ove vizualizacije omogućuju razumijevanje točnosti modela u prepoznavanju različitih klasa. Vidljivo je da model najpreciznije prepoznaje klase Medium i Low, dok se klasa High češće pogrešno klasificira.

Problem pogrešnog klasificiranja visokih vrijednosti rodne nejednakosti može biti posljedica neuravnoteženosti skupa podataka, gdje su instance High klase manje zastupljene u odnosu na druge klase. Kada su klase neravnomjerno raspodijeljene, model može biti skloniji favoriziranju većinskih klasa jer se tijekom obuke susreće s manje primjera High klase. Ovo rezultira nižom preciznošću i odazivom za tu klasu.

Ovakva neuravnoteženost može imati utjecaj na diskriminaciju jer model može precijeniti ili podcijeniti važnost specifičnih skupina podataka. Konkretno, podcijenjene vrijednosti klase High mogu maskirati stvarne slučajeve visoke rodne nejednakosti, što može dovesti do pogrešnih zaključaka o diskriminaciji. Model tako ne prepoznaje kritične uzorke i može nenamjerno ignorirati ili umanjiti važnost skupina koje su podložne većim razinama diskriminacije.

8. Komparativna analiza

Komparativna analiza omogućuje usporedbu specifičnih pokazatelja između različitih entiteta, u ovom slučaju država. Ova tehnika je ključna u procjeni rodne nejednakosti jer pruža uvid u to kako različite zemlje stoje u usporedbi s globalnim prosjecima i prosjecima zemalja s jako visokim HDI. Usporedbom ključnih metrika kao što su indeks rodne nejednakosti, maternalna smrtnost, zastupljenost žena u parlamentu, te obrazovanje i sudjelovanje u radnoj snazi, moguće je identificirati područja s izraženom diskriminacijom.

```
croatia_data = data[data['Country'] == 'Croatia'].iloc[0]
global_averages = data.mean(numeric_only=True)

very_high_hdi_countries = data[data['HUMAN DEVELOPMENT'] == 'VERY HIGH']
high_hdi_key_metrics   =   very_high_hdi_countries[['GII   VALUE',   'Maternal_mortality',
'Adolescent_birth_rate',
                    'Seats_parliamentt(% held by women)', 'F_secondary_educ',
                    'M_secondary_educ', 'F_Labour_force', 'M_Labour_force']]

very_high_hdi_averages = high_hdi_key_metrics.mean()
```

Podaci specifični za Hrvatsku izdvojeni su iz originalnog skupa podataka, a zatim su izračunati globalni prosjeci prosjek država s visokim HDI koristeći funkciju *mean().* Kako bi se osigurala točnost, korišten je parametar *numeric_only=True* u funkciji *mean(),* što omogućuje izračun prosjeka samo za numeričke stupce.

```
comparison_df = pd.DataFrame({
    'Metric': [
        'GII VALUE', 'Maternal_mortality', 'Adolescent_birth_rate',
        'Seats_parliamentt(% held by women)', 'F_secondary_educ',
        'M_secondary_educ', 'F_Labour_force', 'M_Labour_force'
    ],
    'Hrvatska': [
        croatia_data['GII           VALUE'],           croatia_data['Maternal_mortality'],
croatia_data['Adolescent_birth_rate'],
        croatia_data['Seats_parliamentt(%           held            by            women)'],
croatia_data['F_secondary_educ'],
        croatia_data['M_secondary_educ'],                  croatia_data['F_Labour_force'],
croatia_data['M_Labour_force']
    ],
    'Svjetski prosjek': [
        global_averages['GII        VALUE'],        global_averages['Maternal_mortality'],
global_averages['Adolescent_birth_rate'],
        global_averages['Seats_parliamentt(%          held          by          women)'],
global_averages['F_secondary_educ'],
        global_averages['M_secondary_educ'],            global_averages['F_Labour_force'],
global_averages['M_Labour_force']
    ],
    'Prosjek država s visokim HDI': [
        very_high_hdi_averages['GII  VALUE'], very_high_hdi_averages['Maternal_mortality'],
very_high_hdi_averages['Adolescent_birth_rate'],
        very_high_hdi_averages['Seats_parliamentt(%        held        by        women)'],
very_high_hdi_averages['F_secondary_educ'],
        very_high_hdi_averages['M_secondary_educ'],
very_high_hdi_averages['F_Labour_force'], very_high_hdi_averages['M_Labour_force']
    ]
})
```

Kreirani DataFrame *comparison_df* omogućuje lakšu usporedbu između podataka Hrvatske, globalnih prosjeka i prosjek država s visokim HDI. Uključene metrike koje su relevantne za

analizu su: GII iznos, maternalna smrtnost, stopa nataliteta kod adolescentica, postotak žena u parlamentu, te podaci o obrazovanju i radnoj snazi (slika 28) .

	Metric	Hrvatska	Svjetski prosjek	Prosjek država s visokim HDI
0	GII VALUE	0.093	0.339353	0.147032
1	Maternal_mortality	8.000	139.592814	13.983871
2	Adolescent_birth_rate	8.600	43.623952	13.509677
3	Seats_parliamentt(% held by women)	31.100	25.297605	28.395161
4	F_secondary_educ	97.000	62.779641	85.819355
5	M_secondary_educ	100.000	67.069461	87.825806
6	F_Labour_force	45.900	50.218563	52.696774
7	M_Labour_force	58.800	70.088024	69.683871

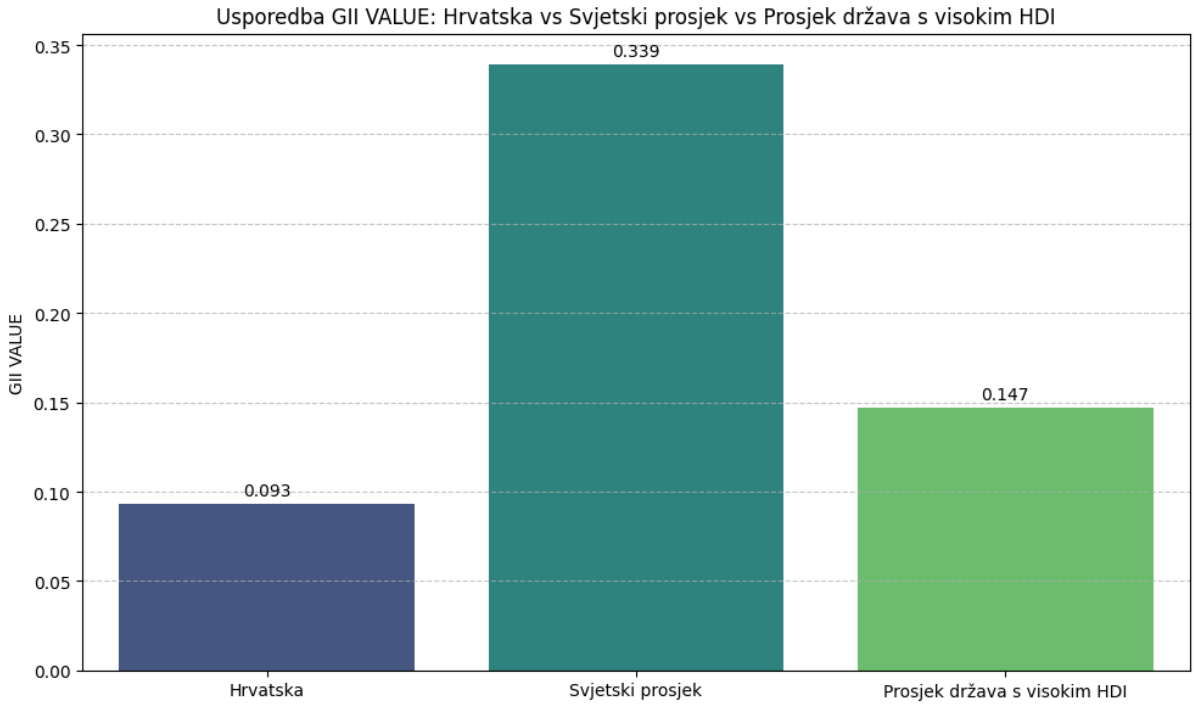
Slika 28. Tablica ključnih metrika za usporedbu

Tablica prikazuje ključne metrike koje omogućuju usporedbu između Hrvatske, globalnog prosjeka i prosjek zemalja s vrlo visokim HDI. Ove metrike su ključne za procjenu rodne nejednakosti i društvene jednakosti u širem smislu.

8.1. Vizualizacije komparativne analize

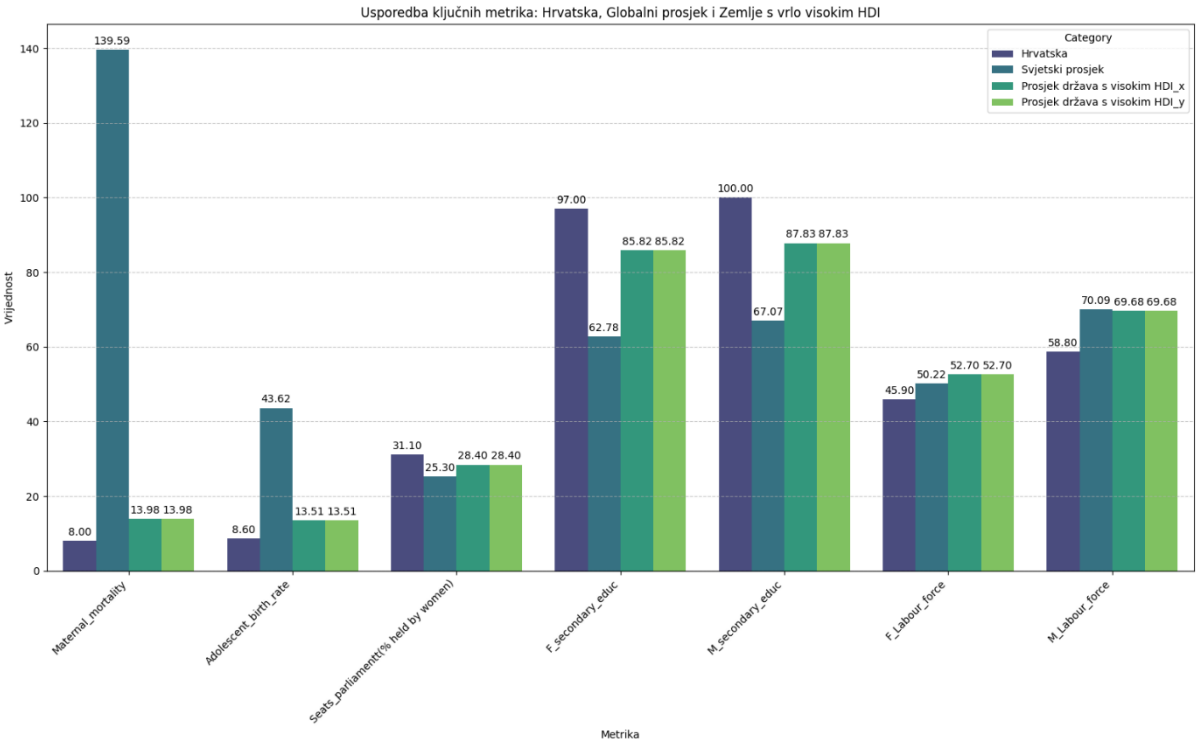
Za vizualizaciju rezultata su odabrane dvije glavne skupine metrika: GII VALUE i ostale metrike. To omogućuje jasnije predstavljanje podataka koji se značajno razlikuju u skalama. Vizualizacija omogućuje jasnije razumijevanje kako se Hrvatska uspoređuje s globalnim standardima i s drugim zemljama visokih razvijenih regija.

GII vrijednost prikazuje rodnu nejednakost, pri čemu niže vrijednosti označavaju nižu rodnu nejednakost. Graf na slici 29 pokazuje da Hrvatska ima značajno nižu GII vrijednost u usporedbi sa svjetskim prosjekom i prosjekom zemalja s vrlo visokim HDI. Ova razlika ukazuje na bolju rodnu ravnotežu u Hrvatskoj u odnosu na globalne i visoko razvijene zemlje, ukazujući na uspješnu politiku usmjerenu prema smanjenju rodnih nejednakosti.



Slika 29. Usporedba GII VALUE: Hrvatska i Svjetski prosjek i Prosjek država s visokim HDI

Vizualizacija ključnih metrika izdvaja Hrvatsku kao pozitivan primjer u usporedbi s globalnim prosjekom i zemljama s vrlo visokim HDI (slika 30). Hrvatska pokazuje iznimno nisku stopu maternalne smrtnosti i stope nataliteta kod adolescentica što ukazuje na uspješan zdravstveni i obrazovni sustav. Postotak žena u parlamentu u Hrvatskoj iznosi 31.10%, što je iznad globalnog prosjeka od 25.30% i prosjeka zemalja s visokim HDI od 28.40%, odražavajući bolju političku participaciju žena. Hrvatska se ističe i u obrazovanju, s 97% žena i 100% muškaraca koji završavaju srednju školu, što je iznad globalnog prosjeka i prosjeka zemalja s visokim HDI. Ipak, stopa sudjelovanja u radnoj snazi za žene (45.90%) i muškarce (58.80%) u Hrvatskoj je ispod prosjeka, što ukazuje na potrebu za dodatnim poboljšanjima u toj oblasti.



Slika 30. Usporedba ključnih metrika: Hrvatska, Globalni prosjek i Zemlje s vrlo visokim HDI

Zaključak

Dubinska analiza podataka pruža moćne alate za otkrivanje skrivenih obrazaca i razumijevanje složenih problema poput rodne diskriminacije, omogućujući donošenje informiranih i ciljnih odluka. U ovom istraživanju primijenjene su raznovrsne tehnike kako bi se detaljno sagledali uzroci i manifestacije rodne nejednakosti.

Deskriptivna analiza otkrila je osnovne značajke i obrasce unutar podataka, dok je regresijska analiza istaknula ključne socio-ekonomske faktore koji pridonose nejednakosti, poput niskog obrazovanja žena i njihove smanjene participacije u politici i radnoj snazi. Za grupiranje zemalja prema sličnim razinama diskriminacije korišteno je klasteriranje, a detekcija anomalija identificirala je zemlje koje značajno odstupaju, naglašavajući kritične slučajeve. Klasifikacija je korištena za predviđanje razina rodne nejednakosti, dok je komparativna analiza omogućila usporedbu ključnih indikatora među državama, pružajući dublji uvid u globalne i lokalne obrasce.

Hrvatska se istaknula u usporedbi s globalnim prosjekom u nekoliko ključnih pokazatelja rodne nejednakosti, uključujući nisku stopu smrtnosti majki, smanjeni natalitet kod adolescentica i visoku razinu završenog srednjoškolskog obrazovanja kod oba spola, što reflektira snažan zdravstveni i obrazovni sustav te bolju političku zastupljenost žena. Međutim, sudjelovanje žena i muškaraca u radnoj snazi ostaje ispod globalnog prosjeka, ukazujući na potrebu za dodatnim mjerama i politikama koje bi potaknule veće zapošljavanje i ravnopravnost na tržištu rada.

Literatura

[1] UNESCO-ov izvještaj o rodnoj nejednakosti u obrazovanju, India Today, Preuzeto 06.08.2024. sa: https://www.indiatoday.in/education-today/gk-current-affairs/story/unescos-report-on-gender-inequality-in-education-314057-2016-03-19

[2] „Data Vs Information Vs Knowledge: Understand The Difference", knowmax.ai, Preuzeto 05.08.2024. sa: https://knowmax.ai/blog/data-vs-information-vs-knowledge/

[3] Tan Pang-Ning, Steinbach Michael, Kumar Vipin, Karpatne Anuj, *Introduction to Data Mining*, Pearson, New York, 2019.

[4] „KDD in Data Mining", scaler.com, Preuzeto 05.08.2024. sa: https://www.scaler.com/topics/data-mining-tutorial/kdd-in-data-mining/

[5] "CRISP-DM", Institut Rudjer Bošković, Preuzeto 07.08.2024. sa: http://dms1.irb.hr/tutorial/hr_dm_proces.php

[6] Fürnkranz Johannes, Gamberger Dragan i Lavrač Nada, *Foundations of Rule Learning*, Springer Science & Business Media, New York, 2012.

[7] Soares Carlos, Peng Yonghong, Meng Jun, Washio Takashi i Zhou Zhi-Hua, *Applications of Data Mining in E-Business and Finance: Introduction*, IOS Press, Nizozemska, 2008.

[8] Witten Ian H., Frank Eibe i Hall Mark A., *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2011.

[9] „Gender Inequality Index (GII) Technical Notes," Human Development Reports, UNDP. Preuzeto 14.08. sa: https://hdr.undp.org/sites/default/files/2021-22_HDR/hdr2021-22_technical_notes.pdf

[10] „A Complete Guide to Linear Regression", kaggle.com, Preuzeto 10.09.2024. sa: https://www.kaggle.com/code/hserdaraltan/a-complete-guide-to-linear-regression

[11] Bishop Crostopher M., *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[12] Han Jiawei, Kamber Micheline i Pei Jian, *Data Mining: Concepts and Techniques,* 3rd ed., Morgan Kaufmann, San Francisco 2011.

[13] Fayyad Usama, Piatetsky-Shapiro Gregory, Smyth Padhraic, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, AAAI 97, Rhode Island, 1996.

[14] Shearer Colin, "The CRISP-DM Model: The New Blueprint for Data Mining", *Journal of Data Warehousing*, The Data Warehousing Institute, Seattle, 2000.

[15] Aggarwal, Charu C, *Data Mining: The Textbook,* Cham: Springer, 2015.

Popis tablica

Popis slika

Prilog 1

Cijeloviti kod:
https://colab.research.google.com/drive/1ytseOwCD8jxQ9PjC2xtwbos5_nPcyr9l?usp=sharing