

Sadržaj

Uvod	1
1. Naslov prvog poglavlja.....	Error! Bookmark not defined.
1.1. Prvo potpoglavljje.....	Error! Bookmark not defined.
1.2. Stilovi za tekst, naslove i podnaslove	Error! Bookmark not defined.
1.3. Stilovi za nabranjanje	Error! Bookmark not defined.
1.3.1. Stilovi za nabranjanje s točkama i crticama ..	Error! Bookmark not defined.
1.4. Slike	Error! Bookmark not defined.
1.5. Tablice	Error! Bookmark not defined.
1.6. Matematički izrazi	Error! Bookmark not defined.
1.7. Programski kôd.....	Error! Bookmark not defined.
Zaključak	37
Literatura	38
Sažetak.....	39
Summary.....	40
Skraćenice.....	Error! Bookmark not defined.
Privitak	Error! Bookmark not defined.

Uvod

Natalitet i mortalitet jedni su od ključnih demografskih pokazatelja trenutnog stanja u državi, ali isto tako i jedan od mogućih pokazatelja budućeg razvitka države. Stoga kvalitetne procjene te predikcije istih mogu uvelike pomoći pri donošenju odluka unutar populacijskih politika, ali i u izgradnji sustava poput obrazovnog, zdravstvenog i mirovinskog.

Razumijevanje pozitivnih te negativnih utjecaja drugih demografskih i ekonomskih varijabli na stope nataliteta i mortaliteta omogućuje nam uvid u promjene koje bi mogle pomoći uspjehu populacijskih politika i napretku države.

Ovo je od posebne važnosti Republici Hrvatskoj kao državi koja se suočava sa niskim natalitetom, visokim mortalitetom te starenjem populacije. Cilj ovog rada je odrediti koje varijable utječu na takve demografske trendove izgradnjom prediktivnih modela.

Postoje već poznati matematički modeli koji se koriste za predikcije populacijskih varijabli poput modela eksponencijalnog i logističkog rasta, Leslie matrice i Gompertz modela. No takvih modeli mogu biti ograničeni pri velikim skupovima podataka te pronalasku nelinearnih odnosa između značajki.

Posljednjih godina upotreba strojnog učenja za izvlačenje znanja iz podataka te rješavanja složenih analitičkih problema sve je raširenije. Također modeli strojnog učenja u mogućnosti su da analiziraju velike količine podataka i prepoznaju složenije odnose.

Upravo takvi prediktivni modeli su fokus ovog rada, u svrhu shvaćanja utjecaja pojedinih demografskih i ekonomskih značajki na stope nataliteta i mortaliteta u Republici Hrvatskoj te predviđanju budućih trendova.

1. Pregled literature i povezanih radova

Budući da su stope nataliteta i mortaliteta jedne od ključnih demografskih značajki svake države, postoje brojni radovi o uzrocima određenih postojećih trendova te predikcijama novih.

Model Lee-Carter[?link] jedan je od najpoznatijih modela za dugoročne prognoze smrtnosti. Zasnovan je na statističkim metodama vremenskih serija te se prilagođava povijesnim podacima. Mnogi radovi zasnovani su upravo na tom modelu. Dodavanjem BDP-a kao varijable, model pruža bolje prilagodbe i bolje interpretativne prognoze (G. Niu, B. Melenberg, 2014. [?link]). Noviji modeli grade se na što širem spektru ekonomskih, okolišnih i životnih čimbenika kako bi pružili veću točnost te otkrili moguće zavisnosti. Takvi modeli superiorniji su od Lee-Carter modela te za pojedine države mogu pružiti točnija predviđanja od Niu-Melenberg modela (Dimai, Matteo, 2023.[?link]).

Stope nataliteta često se povezuju sa ekonomskim čimbenicima i razvojem države. Viši prihodi mogu doprinijeti značajno nižem natalitetu, što znači da mogući ekonomski razvoj države negativno utječe na stope nataliteta (Nick Turner, Kendra Robbins, 2022 [?link]). Jedna od poznatijih metoda za analizu sklapanja brakova je Coale-McNeil metoda(Ansley J. Coale, D. Susan McNeil, 1972. [?link]). Takva metoda može pomoći u predviđanju nataliteta i fertiliteta pošto su usko povezani sa stopama novih brakova.

Posljednjih godina, dominiraju prediktivni modeli temeljeni na strojnom učenju. Modeli linearne regresije te umjetne neuronske mreže u mogućnosti su predvidjeti ponašanja fertiliteta te pružiti znanstvenu osnovu za upravljanje urbanom populacijom (Zhu, Zhu, Gu, Chen, Zhan, Li, Huang, Xu and Li, 2022. [?link]). Za procjene populacije u odsutnosti potpunog i/ili nedavnog popisa stanovništva, može se koristiti Bayesov model koji pruža pouzdane procjene nesigurnosti za podršku informiranim odlukama (Jochem WC, provjeri + [?link]). Algoritmi strojnog učenja, posebice ensemble regresijski modeli poput Extreme Gradient Boosting i CatBoost modela, poboljšavaju predviđanja analiziranjem neizvjesnosti na demografskim podacima te smanjivanjem faktora koji otežavaju predviđanje. Algoritmi strojnog učenja pokazuju zadovoljavajuće rezultate, iako najbolje performanse pružaju na velikim skupovima podataka, što nije moguće kod predviđanja

demografskih čimbenika jer smo ograničeni na male vremenske raspone (Fatih Veli Şahinarslan, Ahmet Tezcan Tekin, Ferhan Çebi, 2021.)

2. Metodologija

2.1. Prikupljanje podataka

Podatci korišteni u ovom radu nabavljeni su iz različitih izvora. Podatci o broju stanovništva, vjenčanih, rođenih i umrlih te harmonizirani indeks potrošačkih cijena prikupljeni su iz baze podataka Eurostata.^(link na literaturu) Većina ostalih demografskih pokazatelja poput očekivanog trajanja života, neto migracija, stanovništva ruralnog, odnosno urbanog područja, stanovništva u najvećem gradu, broj stanovništva u određenom dobnom rasponu, odnosi ovisnosti o dobi, stope smrtnosti i rodности na 1000 ljudi, te ekonomskih pokazatelja tečaja valute, tržišne kapitalizacije dionica u odnosu na BDP i indeksa potrošačkih cijena, pribavljeni su iz baze podataka WorldBanka. ^(link) Podatci o nezaposlenosti, inflaciji, uvozu, izvozu i proizvodnji preuzeti su sa stranice MacroTrends. ^(link) Indeksi neto plaća te indeksi pouzdanja, očekivanja i raspoloženja potrošača su preuzeti sa stranice HNB-a. ^(link) Podatci o BDP-u preuzeti su iz baze podataka Maddison projekta. ^(link)

2.2. Opis podataka

Većina prikupljenih podataka je na godišnjoj razini. Mjesečni podatci, kao što su indeksi plaća, indeksi pouzdanja, očekivanja i raspoloženja potrošača, te harmonizirani indeks potrošačkih cijena, preračunati su u godišnje vrijednosti kao prosjek mjesečnih podataka za svaku godinu.

Podatci su dostupni i korišteni u različitim vremenskim razdobljima.

- Podatci korišteni u razdoblju od 1960. do 2021. godine: broj ukupnog stanovništva, broj rođenih i stope rodности, broj umrlih i stope smrtnosti, stanovništvo urbanog i ruralnog područja, očekivano trajanje života, stanovništva u određenim dobnim rasponima, stanovništvo najvećeg grada, neto migracije, bruto domaći proizvod
- Podatci korišteni u razdoblju od 1986. do 2021. godine: indeksi potrošačkih cijena, tečaj valute, stopa inflacije

Ciljne varijable za koje se rade predikcije su ukupan broj rođenih te ukupan broj umrlih po godini.

U donjoj tablici opisane su varijable koje će se koristiti u nastavku rada.

Naziv	Opis varijable
Godina (eng. <i>year</i>)	Godina na koju se podatci odnose
Ukupno stanovništvo (eng. <i>Population, total</i>)	Broj ukupnog stanovništva Republike Hrvatske
Stope rodnosti (eng. <i>Birth rate</i>)	Stope rodnosti, broj rođenih na 1000 ljudi
Stope umrlih (eng. <i>Death rate</i>)	Stope smrtnosti, broj umrlih na 1000 ljudi
CPI(sezonski) (eng. <i>CPI seasonal</i>)	Indeks potrošačkih cijena, sezonski prilagođena
CPI	Indeks potrošačkih cijena, nije sezonski prilagođen
Urbano stanovništvo (eng. <i>Urban population</i>)	Ukupni broj stanovništva urbanog područja
% urbanog stanovništva	Postotak urbanog stanovništva u odnosu na ukupno stanovništvo
Ruralno stanovništvo (eng. <i>Rural population</i>)	Ukupni broj stanovništva ruralnog područja
% ruralnog stanovništva	Postotak ruralnog stanovništva u odnosu na ukupno stanovništvo
Godišnji rast/promjena značajke X (eng. <i>annual growth/change</i>)	Prikazuje godišnji rast/promjenu neke značajke X u postotcima u odnosu na prijašnju godinu(npr. ruralnog stanovništva, inflacije...)
Preživljavanje do 65. godine, muškarci/žene (eng. <i>Survival to age 65, male/female</i>)	Preživljavanje do 65. godine života kao postotak od kohorte novorođenčadi koja bi doživjela 65. godinu života, ako bi bila podvrgnuta stopama smrtnosti prema dobi za određenu godinu
Očekivano trajanje života (eng. <i>Life expectancy</i>)	Očekivano trajanje života u godinama

Omjer ovisnost o dobi, stari/mladi (eng. <i>Age dependency ratio, old/young</i>)	Odnos starije/mlađe populacije u odnosu na radno sposobnu populaciju
Stanovništvo(A-B) (eng. <i>Population ages A-B</i>)	Postotak stanovništva u dobnom rasponu od A do B u odnosu na ukupno stanovništvo
X stanovništvo(A-B)	Postotak X stanovništva u dobnom rasponu od A do B u odnosu na ukupno X stanovništvo, X = {muškog (eng. <i>male</i>), ženskog (eng. <i>female</i>)}
Stanovništvo najvećeg grada (eng. <i>Population in largest city</i>)	Postotak stanovništva u najvećem gradu u odnosu na urbano stanovništvo
Neto migracije (eng. <i>Neto migration</i>)	Neto migracije, negativan ako je broj iseljenih veći od useljenih
BDP (eng. <i>GDP – Gross Domestic Product</i>)	Bruto domaći proizvod po stanovniku, u odnosu na cijene 2011. godine (eng. <i>GDP per capita, 2011 prices</i>)
Tečaj valute (eng. <i>Exchange rate</i>)	Tečaj nove lokalne valute prema američkom dolaru, proširen unatrag, prosječna vrijednost razdoblja
Stopa inflacije (eng. <i>Inflation rate</i>)	Stope inflacije u postocima
Stope nezaposlenih (eng. <i>Unemployment rate</i>)	Postotak nezaposlenih u odnosu na ukupno aktivno stanovništvo
Apsolutni uvoz (eng. <i>Imports-Billions of US \$</i>)	Vrijednost uvoza izraženu u milijardama američkih dolara
Uvoz u % (eng. <i>% of GDP - Imports</i>)	Postotak BDP-a koji predstavlja uvoz
Apsolutni izvoz (eng. <i>Exports-Billions of US \$</i>)	Vrijednost izvoza izraženu u milijardama američkih dolara
Izvoz u % (eng. <i>% of GDP - Exports</i>)	Postotak BDP-a koji predstavlja izvoz
Proizvodnja u % (eng. <i>% of</i>	Postotak BDP-a koji predstavlja proizvodnju

2.3. Analiza podataka

Analiza podataka uključivala je provjeru koreliranosti na standardiziranim podacima između različitih varijabli.

Koreliranost se mjeri koeficijentom korelacije, koji se definira kao statistička mjera snage linearne povezanosti između dviju varijabli. Koeficijent korelacije je definiran u području od -1 do 1. Negativan iznos predstavlja negativnu, odnosno inverznu korelaciju između varijabli. Tada su te dvije varijable obrnuto proporcionalne, kada jedna raste, druga pada. Pozitivan iznos predstavlja pozitivnu korelaciju između varijabli. Tada su varijable proporcionalne te kada jedna raste, raste i druga. Apsolutne vrijednosti tih iznosa bliže nuli predstavljaju manji stupanj linearne povezanosti, dok apsolutni iznosi bliži 1 predstavljaju jači stupanj linearne povezanosti. Korelacijski koeficijent iznosa 0 znači da nema linearne povezanosti, iznos -1 znači savršenu negativnu, a 1 savršenu pozitivnu povezanost.

Detaljno opisan postupak analize te rezultati nalaze se u poglavlju „Primjena metodologije“ [[?link](#)].

2.4. Modeli

Cilj ovog rada je izrada prediktivnog modela za predviđanje nataliteta i mortaliteta. U ovom potpoglavlju predstavljeni su svi korišteni modeli, objašnjen je njihov princip rada i arhitektura.

Svi modeli su učeni u nadziranom okruženju, odnosno za svaki ulazni skup podataka za učenje poznate su izlazne vrijednosti ciljne varijable. Svi podatci su numeričkog tipa te nije bila potrebna klasifikacija. Izlaz se sastoji od jedne ciljne varijable. Modeli su pisani u programskom jeziku python.

Vrednovanje naučenih modela provođeno je primjenom sljedećih evaluacijskih metrika:

- Srednja apsolutna pogreška (eng. *Mean Absolute Error*, MAE) – prosjek apsolutnih pogrešaka između stvarnih vrijednosti i predikcija

- Srednja kvadratna pogreška (eng. *Mean Squared Error*, MSE) – prosjek kvadrata pogrešaka između predviđenih i stvarnih vrijednosti
- Korijen srednje kvadratne pogreške (eng. *Root Mean Squared Error*, RMSE) – korijen kvadrata srednje kvadratne pogreške, slično kao MAE, ali više kažnjava veće pogreške
- Srednja apsolutna postotna pogreška (eng. *Mean Absolute Percentage Error*, MAPE) – prosječna apsolutna pogreška kao postotak stvarnih vrijednosti, mjeri koliko su predikcije u prosjeku odstupale od stvarnih vrijednosti,

Manje vrijednosti MAE, MSE, RMSE i MAPE znače bolje procjene te točniji model. Ne postoje točno određeni iznosi za koje bi značilo da je model dobar te vrijednosti MAE, MSE i RMSE uvelike ovise o podacima. Naprimjer kod podataka koji su u milijunima MAE od nekoliko stotina može biti zanemariva pogreška, no kod podataka koji su u stotinama, pogreške od stotinu zapravo znače da model ima iznimno visok stupanj netočnosti. Vrijednosti do 10% za MAPE se generalno smatraju zadovoljavajućima. Važno je napomenuti da se nijedna od ovih mjera ne može samostalno smatrati dovoljnom za potpunu procjenu modela te zadovoljavajući rasponi mjera ovise o podacima, domeni te zahtjevima točnosti.

Važan pojam u strojnom učenju koji će se spominjati u analizi modela je prenaučенost (eng. *overfitting*). Do prenaučенosti dolazi kada se model uči na šumovima i/ili nepotrebnim podacima te ostvaruje odlične performanse na podacima za učenje, ali na novim neviđenim podacima ima loše performanse. Prenaučenost će se provjeravati usporedbom performansi na podacima za učenje i podacima za testiranje te analizom krivulje učenja (eng. *learning curve*) tijekom iteracija modela.

2.4.1. Linearna regresija

Linearna regresija je algoritam koji pruža linearnu ovisnost između varijabli kako bi predvidjeli buduće trendove. Koristi se kao statistička metoda u znanosti o podacima te strojnom učenju. Linearna regresija se koristi kada je izlaz neka kontinuirana ili brojčana vrijednost.

Linearna regresija pokušava izračunati izlaznu varijablu kao linearnu kombinaciju ulaznih varijabli. Stoga učenje linearne regresije se svodi na utvrđivanje koeficijenata linearne

jednadžbe iz koje dobivamo izlaze. Struktura linearne regresije sastoji se od ulaza, koeficijenata, jednadžbe i izlaza.

2.4.2. Eksponencijalno izgladivanje i Holtova metoda

Eksponencijalno izgladivanje (eng. *Exponential smoothing*) je metoda za predviđanje vremenskih serija koja eksponencijalno smanjuje važnost starijih podataka, odnosno podatci bliži trenutnom imaju veću težinu. Ova metoda također “izgladjuje” podatke kako bi smanjila buku (odstupanja i krive podatke) te otkrila osnovni trend u vremenskim serijama. Postoje tri vrste eksponencijalnog izgladivanja: jednostavno, dvostruko te trostruko. Jednostruko eksponencijalno zagladivanje pretpostavlja da ne postoji nikakav trend. Dvostruko eksponencijalno zagladivanje, još poznato kao Holtovo linearno eksponencijalno zagladivanje, odnosno Holtova metoda koristi dva parametra za izgladivanje. Holtova metoda se koristi kod podataka koji imaju trend. Trostruko eksponencijalno zagladivanje se još naziva i Holt-Winters metoda. Ono koristi tri parametra za izgladivanje te pomaže u uočavanju sezonskih trendova.

2.4.3. ARIMAX

ARIMA je akronim za autoregresivni integrirani pokretni prosjek (eng. *Autoregressive Integrated Moving Average*). Model ARIMA je jedan od najčešće korištenih modela za predviđanje vremenskih serija. Autoregresivni znači da model koristi prethodne podatke za predviđanje budućih vrijednosti. Integrirani dio predstavlja broj diferenciranja, potreban da bi vremenska serija bila stacionarna. Stacionarne vremenske serije imaju konstantnu srednju vrijednost, varijancu i kovarijancu kroz vrijeme, što omogućava lakše predviđanje uzoraka. Posljednji dio modela je pokretni prosjek koji koristi prijašnje greške modela kao jednu od informacija za točnije predviđanje. Svaki od ovih dijelova ARIMA modela ima svoj parametar. Parametar p mjeri korelaciju između trenutne vrijednosti vremenske serije i prijašnjih podataka vremenske serije te pripada autoregresivnom dijelu modela. Parametar d predstavlja integrirani dio, odnosno broj diferenciranja. Parametar q mjeri korelaciju između trenutne vrijednosti vremenske serije i podataka o greškama na prijašnjim predviđanjima te predstavlja pokretni prosjek. ARIMAX proširuje ARIMA model dodavanjem egzogenih varijabli, odnosno omogućuje nam izradu modela upotrebljavajući više varijabli kao dodatne faktore u predviđanju novih vrijednosti vremenske serije.

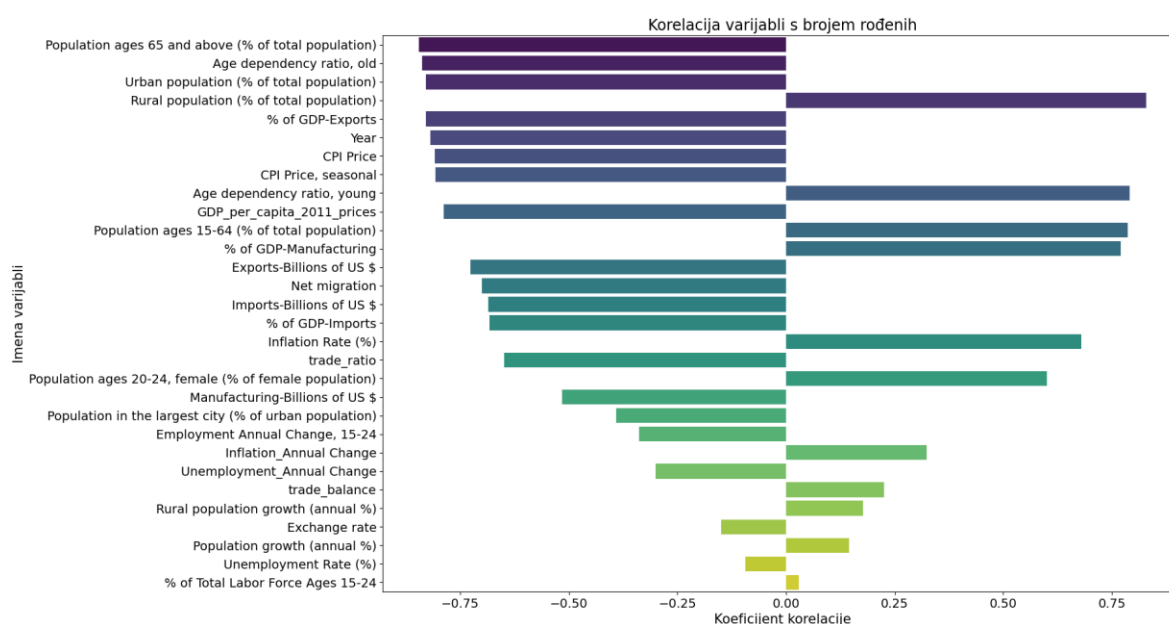
2.4.4. Extreme gradient boosting

Extreme Gradient Boosting (XGBoosting) algoritam je strojnog učenja koji radi na temelju „gradient boosting“ stabala odluke. „Gradient boosting“ stabla koriste tehniku „boostinga“ za izgradnju modela, odnosno kombiniraju više jednostavnih modela nazivanih „slabi učitelji“ kako bi se stvorio bolji, precizniji model. Algoritam započinje treniranjem jednostavnog modela. Pogreške tog modela koriste se u funkciji gubitka za izgradnju sljedećeg modela koji pokušava ispraviti pogreške trenutnog. Novi model se kombinira sa početnim kako bi poboljšao njegove predikcije. Ti koraci se ponavljaju zadanim brojem puta. Jedno od glavnih svojstava ovog algoritma je sprečavanje prenaučivosti. To omogućuje funkcija cilja, koja se sastoji od gore navedene funkcije gubitka te funkcije regularizacije koja kontrolira složenost modela. XGBoosting ima široku uporabu jer je u mogućnosti raditi s velikim skupovima podataka te se može koristiti i u problemima klasifikacije i u problemima regresije.

3. Primjena metodologije

3.1. Koreliranost značajki sa brojem rođenih

Izračunom koreliranosti između broja rođenih sa svakom od ostalih varijabli dobiven je sljedeći graf:



Na x-osi su prikazani koeficijenti korelacije, a na y-osi imena varijabli.

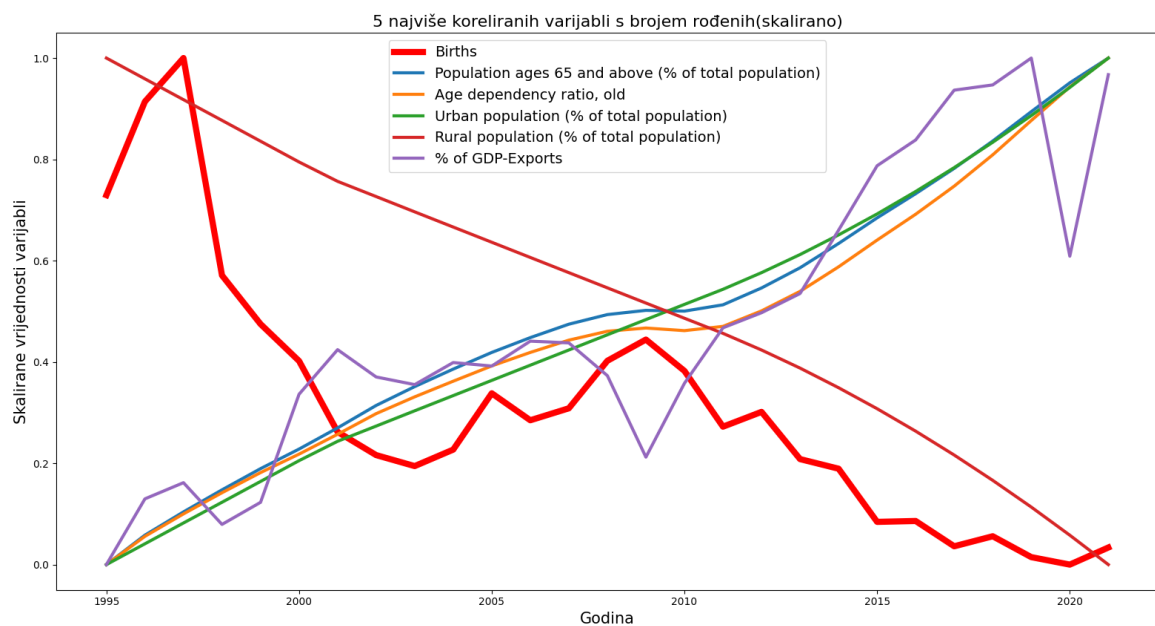
Sa grafa možemo iščitati neke očekivane odnose:

- udio starog stanovništva negativno koreliran s brojem rođenih te udio mlađeg stanovništva pozitivno koreliran s brojem rođenih
- udio urbanog stanovništva negativno koreliran s brojem rođenih
- udio stanovništva u ruralnom području pozitivno koreliran s brojem rođenih
- indeks potrošačkih cijena negativno koreliran s brojem rođenih
- veći udio proizvodnje u BDP-u pozitivno koreliran s brojem rođenih
- udio mladih žena pozitivno koreliran s brojem rođenih
- postotak proizvodnje u BDP-u pozitivno koreliran s brojem rođenih

Zanimljiviji neočekivani odnosi:

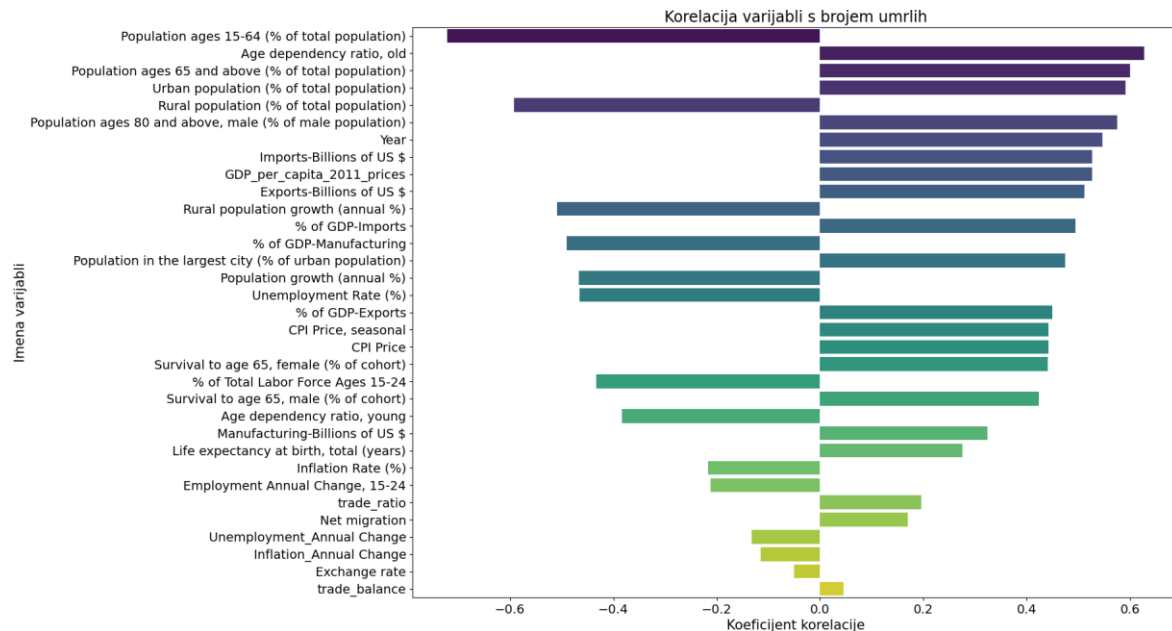
- BDP negativno koreliran s brojem rođenih
 - Veći BDP može značiti mogući veći razvoj države, urbanizaciju, veću uključenost žena na tržištu rada, više obrazovanje, veću dostupnost kontracepcije što može negativno utjecati na broj rođenih
- Postotak izvoza u BDP-u negativno koreliran s brojem rođenih
 - Isto kao i za BDP te moguća globalizacija kojom se mijenjaju društvene vrijednosti i percepcija obitelji
 - Veća ovisnost o vanjskom tržištu može dovesti do neizvjesnosti o radnom mjestu
- Stopa inflacije pozitivno korelirana s brojem rođenih
 - Veća inflacija može značiti gospodarski rast
 - Budući da za stope inflacije imamo dostupno razdoblje tek od 1986. moguće su velike fluktuacije zbog rata 90-tih te ekonomske krize 2008. što dovodi do mogućih krivih odnosa koreliranosti

Na sljedećem grafu možemo vidjeti kako se vrijednosti 5 najviše koreliranih varijabli i broj rođenih mijenja kroz vrijeme.



3.2. Koreliranost značajki sa brojem umrlih

Izračunom koreliranosti između broja umrlih sa svakom od ostalih varijabli dobiven je sljedeći graf:



Kao i kod prijašnjeg grafa, x-os označava koeficijent korelacije, dok y-os označava varijable.

Na grafu možemo vidjeti očekivane odnose:

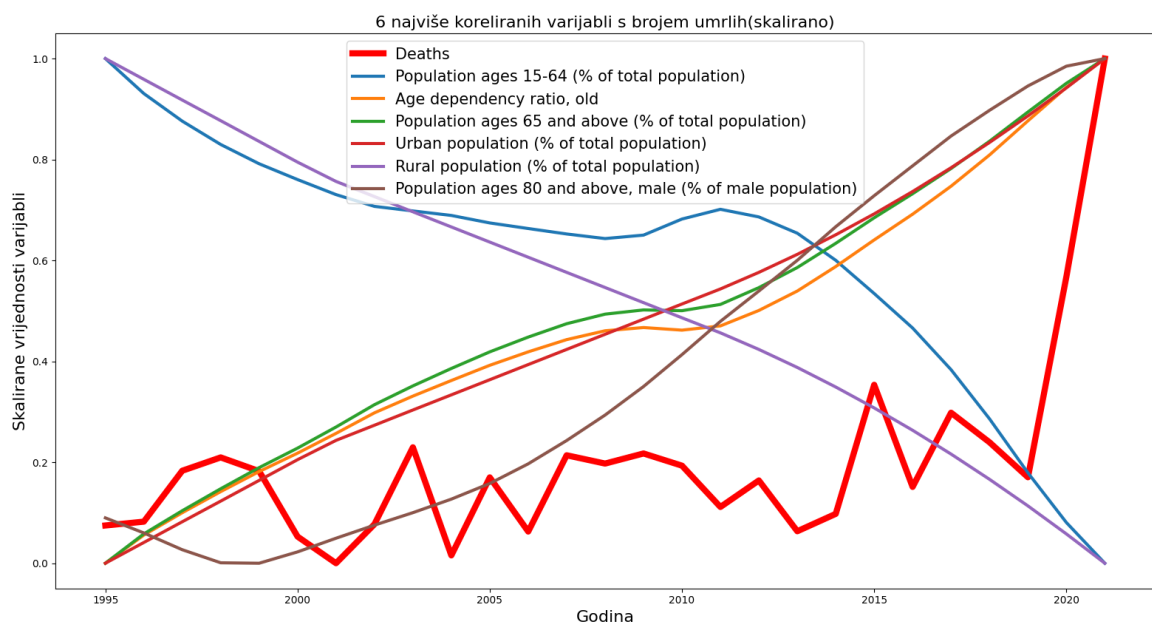
- Pozitivna koreliranost godine s brojem umrlih
- Rast populacije negativno koreliran s brojem umrlih
- Udio starijeg stanovništva pozitivno koreliran s brojem umrlih te udio mlađeg stanovništva negativno koreliran s brojem umrlih

Manje očekivani odnosi:

- Postotak urbane populacije pozitivno koreliran s brojem umrlih, a postotak ruralne populacije negativno koreliran s brojem umrlih
 - Iako veći udio urbane populacije može značiti bolju zdravstvenu skrb to u Republici Hrvatskoj nije nužno slučaj, također do većeg udjela urbane populacije dolazi kod razvijenijih država koje se bore sa starenjem stanovništva i većom smrtnošću

- Iako bi veći udio ruralnog stanovništva mogao značiti lošiju zdravstvenu skrb i kraći životni vijek, kao i kod gornje crtice o urbanoj populaciji, takav trend možemo pripisati sveukupnom razvoju države
- BDP i izvoz pozitivno korelirani s brojem umrlih, udio nezaposlenih negativno koreliran s brojem umrlih
 - Veći BDP i izvoz te manji udio nezaposlenih mogu označavati veći razvoj države, što opet može dovesti do starenja stanovništva odnosno većeg broja umrlih

Na sljedećem grafu možemo vidjeti promjene broja umrlih te odabranih koreliranih varijabli kroz vrijeme.



3.3. Linearna regresija

Za izradu modela korišteni su sljedeći uvozi:

- `LinearRegression` iz `sklearn.linear_model` biblioteke za izradu modela
- `Train_test_split` iz `sklearn.model_selection` biblioteke za procjenu uspješnosti modela

Svi modeli su koristili 70% podataka za treniranje te 30% podataka za testiranje.

3.3.1. Linearna regresija za broj rođenih

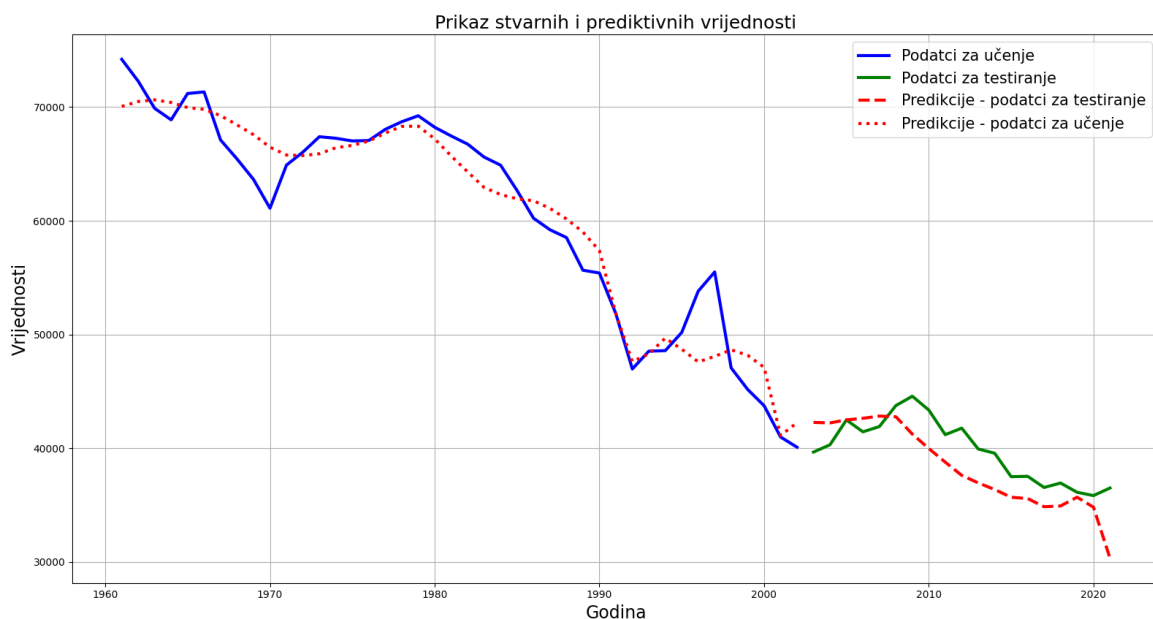
Model linearne regresije za broj rođenih s najboljim performansama za podatke u razdoblju od 1960. do 2022. godine uključivao je sljedeće varijable:

- Godina
- Stanovništvo najvećeg grada (% urbanog stanovništva)
- Godišnji % rast ruralnog stanovništva
- Stanovništvo (15-64)
- Omjer ovisnosti o dobi, mladi
- BDP po stanovniku

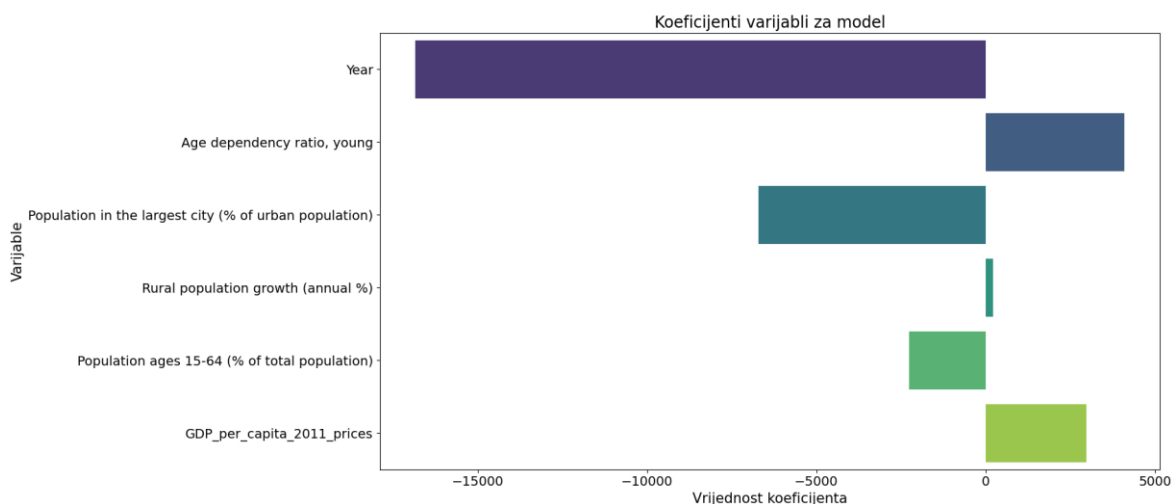
Model ima sljedeće performanse na neviđenim podacima:

- MAE: 2219.53
- MSE: 6927244.17
- RMSE: 2631.97
- MAPE: 0.056

Budući da je broj rođenih u rasponu od oko 30,000 do 80,000, MAE od 2200 nije previše značajan, a MAPE od 5,6% smatramo zadovoljavajućim. No, model nije savršen te na donjem grafikonu možemo vidjeti da, iako je uspio uhvatiti opći trend, postoje odstupanja u predviđanjima, posebice za 2021 godinu.



Važan dio modela su varijable i kako one na njega utječu. Na sljedećem grafu prikazani su koeficijenti modela za svaku ulaznu varijablu.



Ako bi uklonili varijablu “Year“ model ima potpuno krive predikcije. Može se popraviti uključivanjem novih varijabli, no nije moguće postići performanse gornjeg modela.

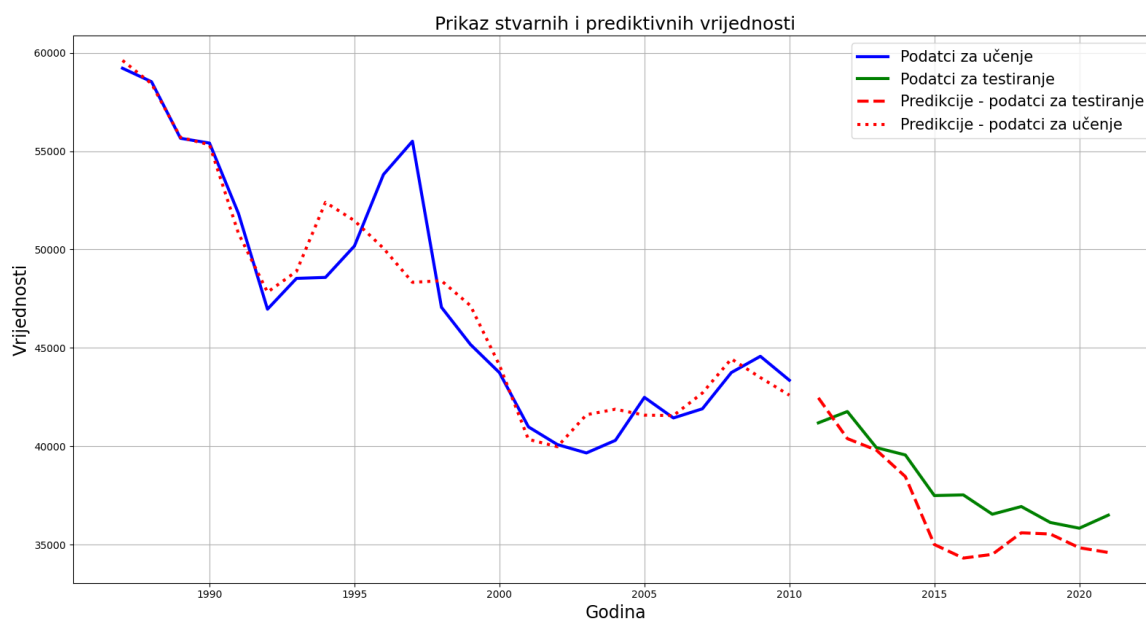
Model linearne regresije za broj rođenih sa podacima u razdoblju od 1986. do 2022. godine uključivao je sve varijable prijašnjeg modela te dodatno:

- CPI, sezonski
- Tečaj valute

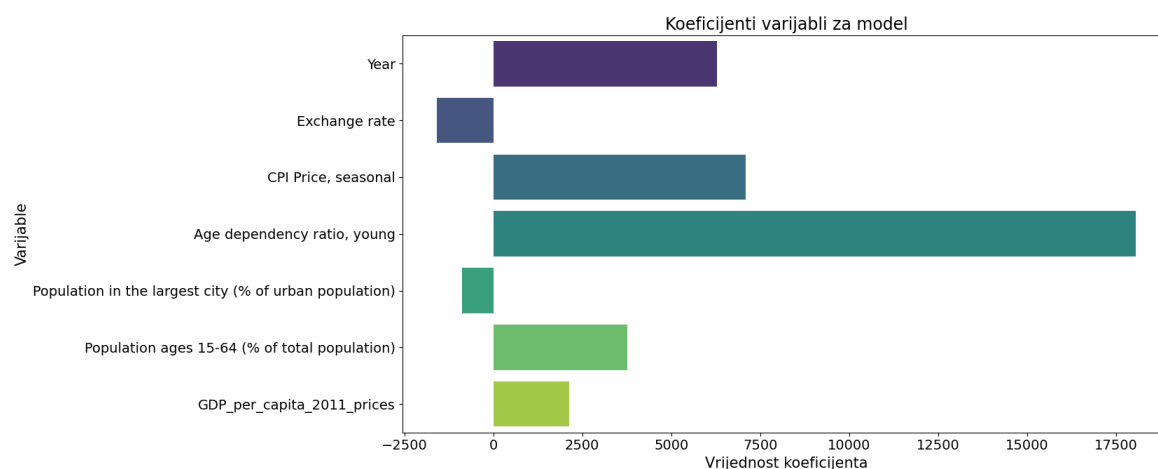
Model je postigao bolje rezultate od prijašnjeg te ima sljedeće performanse na neviđenim podacima:

- MAE: 1492.09
- MSE: 2915605.69
- RMSE: 1707.51
- MAPE: 0.0395

Na donjem grafu možemo uočiti da predviđanje ovog modela za 2021. godinu nema toliko veliko odstupanje kao prijašnji model.



Model na vremenskom razdoblju od 1986. do 2022. godine, ali bez novih varijabli ima znatno lošije performanse. Stoga možemo zaključiti da je uvođenje dviju dodatnih novih varijabli dovelo do poboljšanja performansi. Tu ovisnost možemo vidjeti i na donjem grafu koeficijenata varijabli modela.



Također možemo uočiti da varijabla “Age dependency ratio, young“ u ovom modelu ima puno veću važnost u odnosu na prethodni. Takvi rezultati mogući su zbog fluktuacija u starijim podacima.

3.3.2. Linearna regresija za broj umrlih

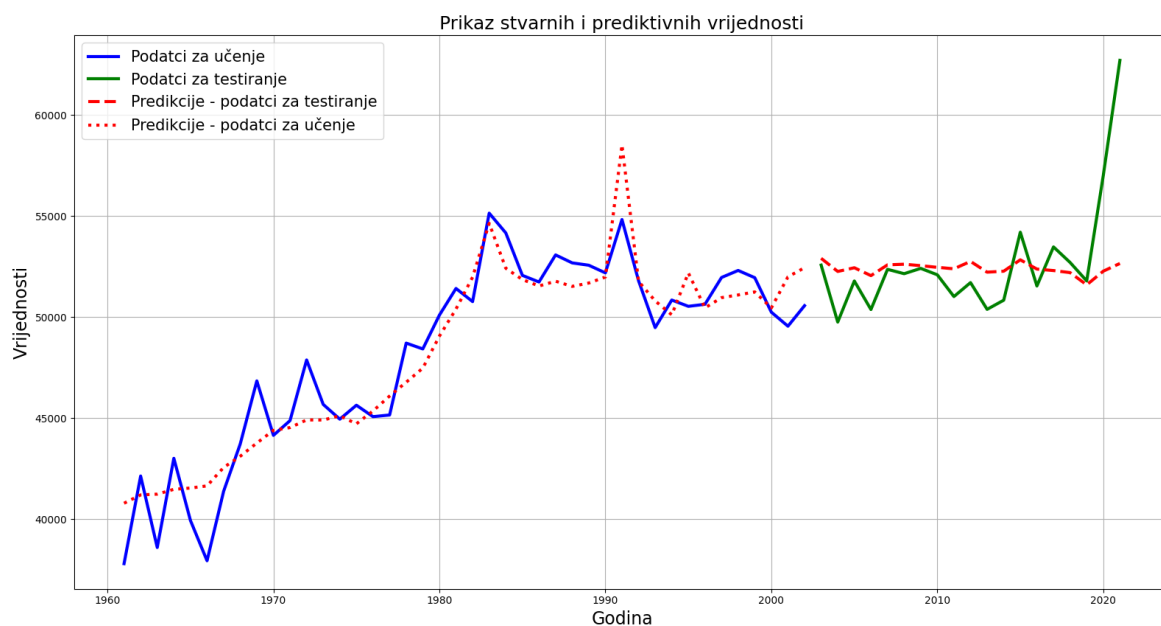
Model linearne regresije za broj umrlih s najboljim performansama za podatke u razdoblju od 1960. do 2022. godine uključivao je sljedeće varijable:

- Godina
- Omjer ovisnosti o dobi, stari
- Očekivano trajanje života
- Preživljavanje do 65. godine, muškarci

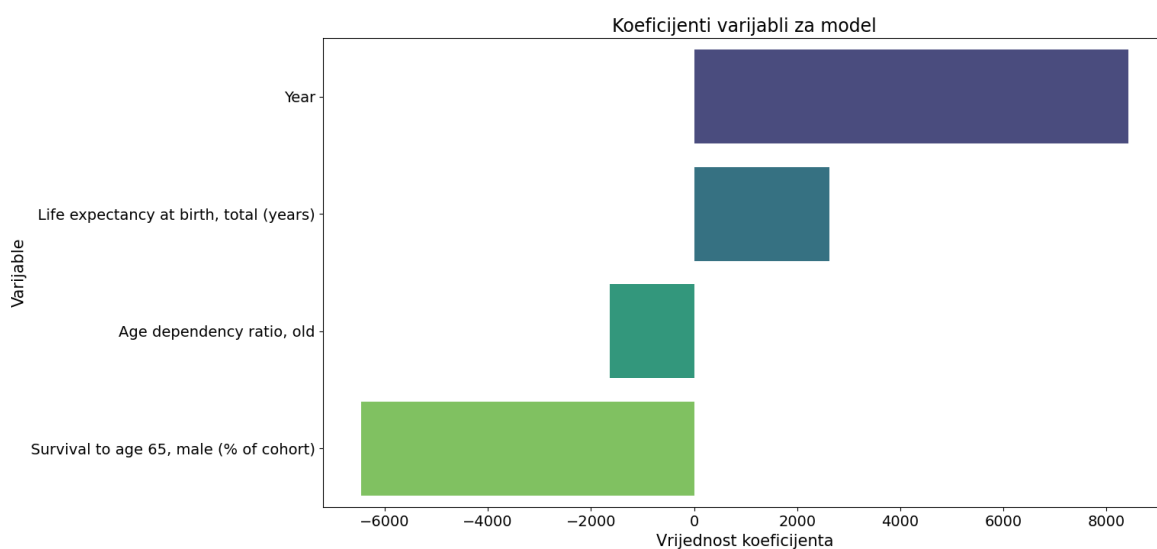
Model ima sljedeće performanse na neviđenim podacima:

- MAE: 1628.52
- MSE: 7702720.98
- RMSE: 2775.38
- MAPE: 0.0294

Na temelju gornjih vrijednosti možemo zaključiti da model ima zadovoljavajuće rezultate, te isto možemo vidjeti na donjem grafu stvarnih i prediktivnih vrijednosti. Jedino značajno odstupanje je ponovno za 2021. godinu.



Model najviše ovisi o varijablama “Year“ i “Survival to age 65“ što možemo vidjeti na donjem grafu.



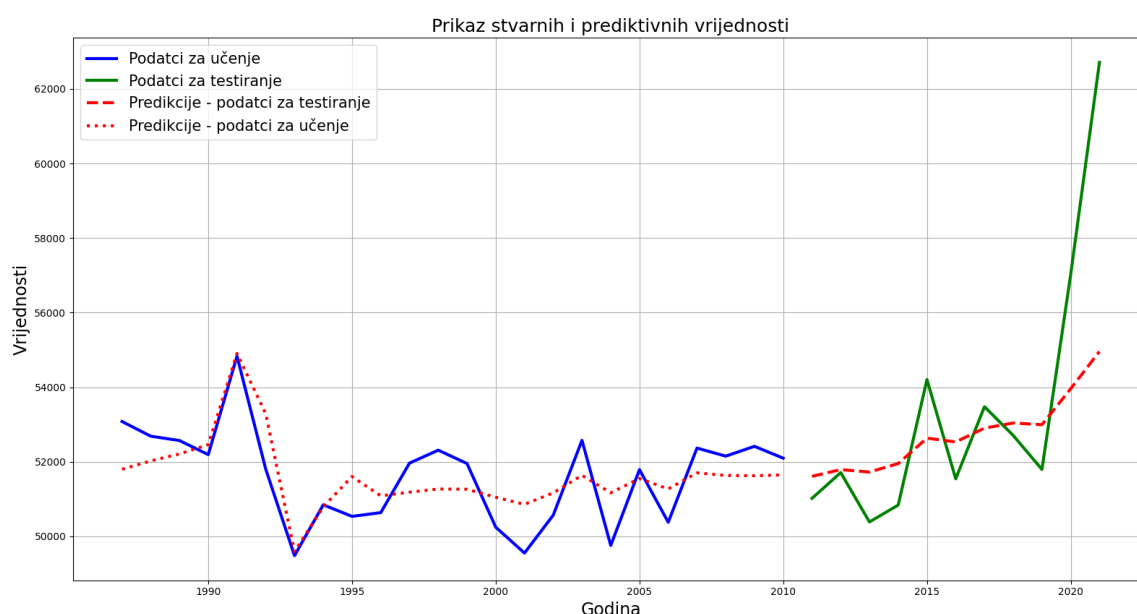
Model za vremenski period od 1986. do 2021. godine ima malo lošije performanse od gornjeg modela, no još uvijek zadovoljavajuće. Uklonjena je varijabla “Age dependency ratio“ te su dodane sljedeće varijable:

- CPI
- CPI, sezonski

Model najviše ovisi o varijabli „Godina“, a o ostalim podjednako te se kod uklanjanja jedne od njih zanemarivo smanji točnost modela. Performanse modela su sljedeće:

- MAE: 1692.02
- MSE: 8498970.49
- RMSE: 2915.3
- MAPE: 0.0298

Bez dodavanja ekonomskih varijabli iznos MAPE modela raste na 0.0453, no i dalje ima zadovoljavajuće performanse. Na donjem grafu možemo vidjeti da model i dalje prati trend te ponovno odstupa za 2021. godinu.



3.4. Holt i Holt-Winters modeli

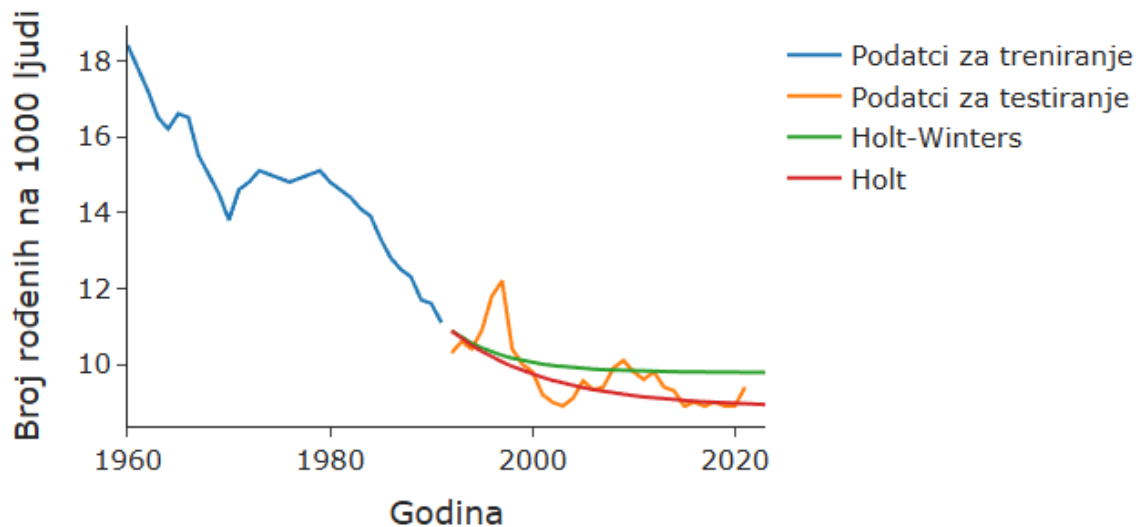
Kod modela eksponencijalnog izgladivanja kao ulaz se koriste samo prijašnji podatci ciljne varijable. Stoga u ovim modelima imamo samo jednu varijablu, stope broja rođenih odnosno umrlih.

Modeli su izgrađeni koristeći 'Holt' za Holtov model te 'ExponentialSmoothing' za Holt-Winters model iz biblioteke 'statsmodels.tsa.holtwinters'.

3.4.1. Modeli za stope rođenih

Predikcije modela za stope rođenih su prikazane na donjem grafu.

Modeli s izглаđivanjem



Možemo primjetiti da Holtov model prati trend bolje, no niti jedan od modela nema značajna odstupanja. To možemo potvrditi i sa sljedećim performansama modela:

Holt-Winters:

- MAE: 0.58
- MSE: 0.53
- RMSE: 0.73
- MAPE: 0.0596

Holt:

- MAE: 0.43
- MSE: 0.39
- RMSE: 0.63
- MAPE: 0.0415

Kod oba modela korišten je parametar 'damped_mod = True' što predstavlja prigušenje. Kada je parametar postavljen na 'False', model će zauvijek pratiti trend, odnosno nastaviti će konstantno rasti ili padati u beskonačnost. S parametrom postavljenim na 'True' model će se nakon nekog vremena prigušiti trend na ravnu liniju. Iz podataka o modelima iščitano

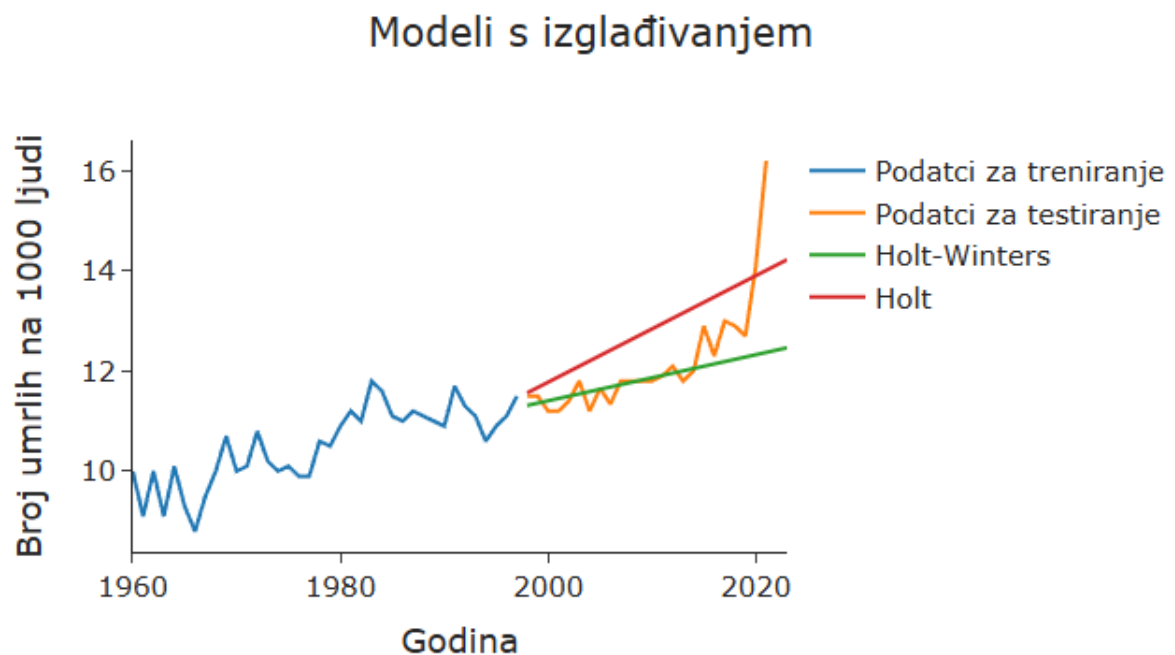
je 'damped_trend' parametar iznosi 0.9 za Holtov model te 0.82 za Hol-Winters model. Takvi iznosi ukazuju na snažno prigušenje trenda, odnosno utjecaj prijašnjih promjena trenda brzo će se smanjivati sa novim predviđanjima.

Parametar 'smoothing_trend' je oko 0.2 za oba modela što sugerira da model sporije reagira na promjene u trendu, što može pomoći kod odstupanja.

Kod oba modela, parametar 'smoothing_level' jednak je 1. Takav iznos pokazuje da modeli brzo reagiraju na promjene u osnovnoj razini jer koriste samo posljednju vrijednost.

3.4.2. Modeli za stope umrlih

Predikcije modela za stope umrlih su prikazane na donjem grafu. Možemo vidjeti da Holt-Winters ima bolje performanse te prati ukupni trend s manjim odstupanjima.



Točne performanse modela su:

Holt-Winters

- MAE: 0.46
- MSE: 0.85
- RMSE: 0.92
- MAPE: 0.0333

Holt

- MAE: 0.82
- MSE: 0.88
- RMSE: 0.94
- MAPE: 0.0671

Kod ovih modela, bolje performanse su ostvarene bez prigušenja trenda, odnosno 'damped_mod' je postavljen na 'False'. Kod Holt modela 'smoothing_level' jednak je 0.9, a 'smoothing_trend' 0.25. Takvi parametri označavaju veliki utjecaj neposredno prethodnih iznosa vremenske serije te mali utjecaj promjena u trendu. Hol-Winters model ima parametre 'smoothing_level' 0.35 te 'smoothing_trend' skoro jednak nuli, što ukazuje na umjereno zaglađivanje razine i vrlo sporo prilagođavanje promjenama u trendu.

3.5. ARIMAX

Za potrebe izgradnje ARIMAX modela korišten je uvoz modela 'SARIMAX' iz 'statsmodels.tsa.statespace.sarimax' biblioteke te su parametri za sezonski utjecaj postavljeni na 0.

Svi modeli su trenirani na 60% podataka te testirani na preostalih 40%.

3.5.1. ARIMAX za broj rođenih

Za podatke od 1960. do 2021. godine, model pokazuje najbolje performanse sa sljedećim varijablama:

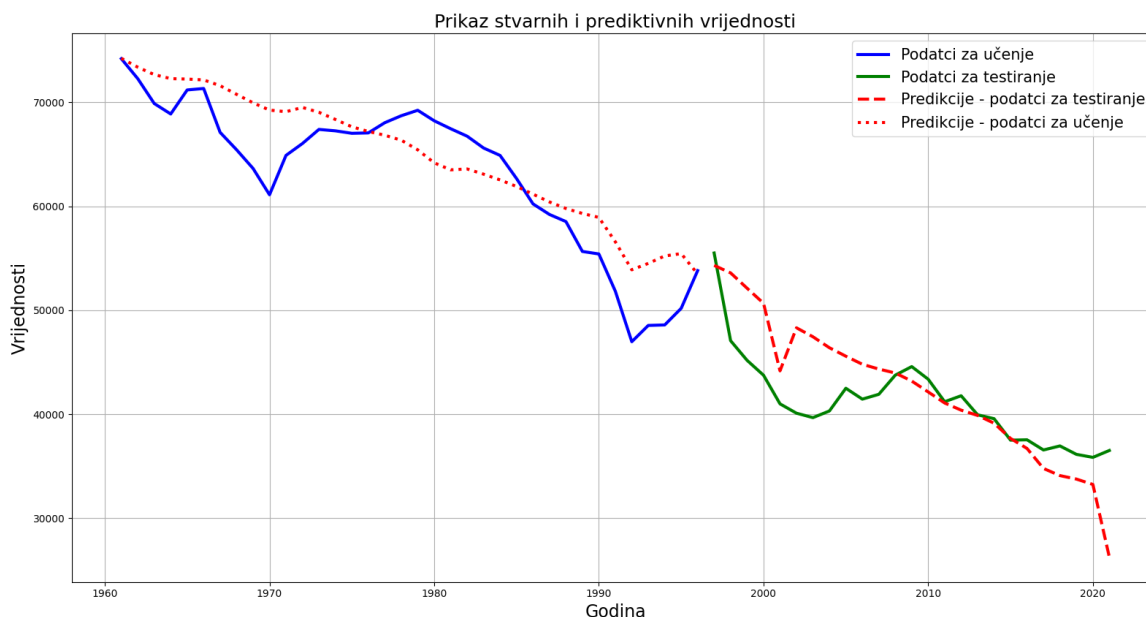
- Godina
- Neto migracije
- Godišnji rast ruralnog stanovništva
- Stanovništvo najvećeg grada

Model ima sljedeće performanse na neviđenim podacima:

- MAE: 2983.47
- MSE: 15989577.24

- RMSE: 3998.7
- MAPE: 0.0719

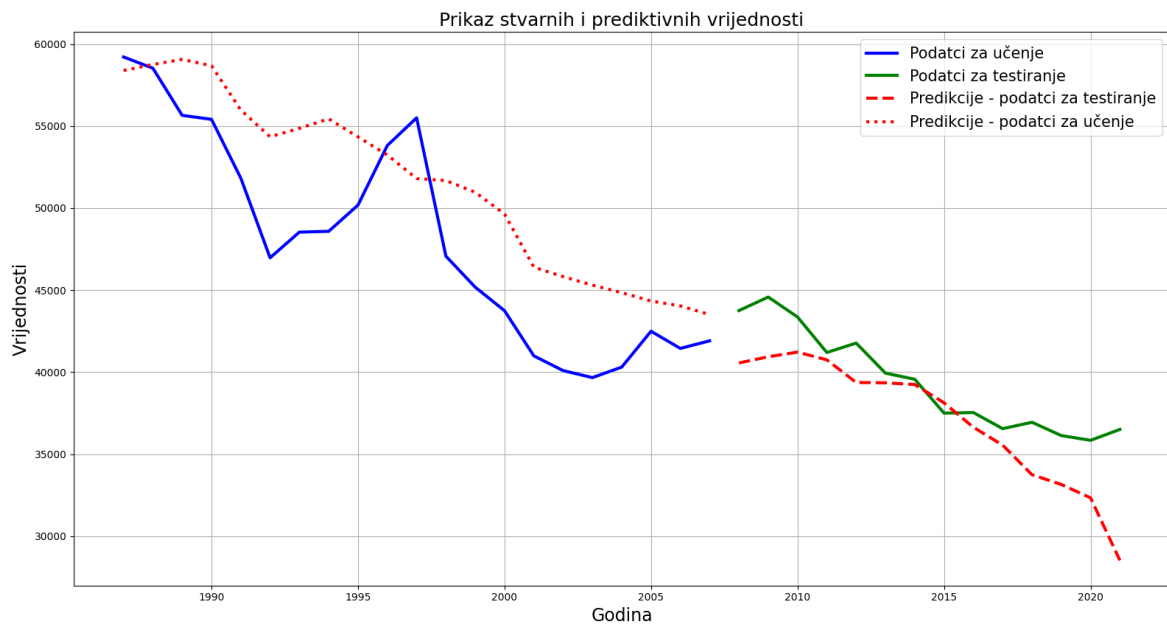
MAE od skoro 3000 ukazuje na osjetna odstupanja od stvarnih vrijednosti. Unatoč tome, model je zadovoljavajući, što potvrđuje MAPE od 7.19%. To znači da model uočava trend, ali ne savršeno što možemo vidjeti i na sljedećem grafu. Također uočavamo da model ima najveće odstupanje za 2021. godinu.



Model najviše ovisi o varijabli „Godina“ te uklanjanjem iste dovodi do jako velikih odstupanja. Dodavanje novih varijabli modelu bez "godine" poboljšava performanse, ali MAPE uvijek ostaje iznad 10%.

Utjecaji AR (autoregresivnih) te MA(pokretnog prosjeka) parametara na model su vrlo mali dok je varijanca greške vrlo visoka i statistički značajna. To ukazuje na nesigurnost modela i sugerira da možda nije optimalan za analizirane podatke.

Za vremensko razdoblje od 1986. do 2021. godine model ne koristi varijable “neto migracije“. Model ima malo bolji MAE i MAPE, što ukazuje na moguće fluktuacije u podacima od 1960. do 1986. koje su dovele do smanjenih performansi iako je skup podataka veći. Predviđanja tog modela možemo vidjeti na sljedećem grafu:



Možemo primijetiti veća odstupanja za godine neposredno prije 2021. godine te ponovno za samu 2021. godinu. Veća odstupanja se mogu vidjeti i na predviđanjima na podacima za učenje. Kao i kod prijašnjeg modela, utjecati AR i MA parametara su zanemarivi.

3.5.2. ARIMAX za broj umrlih

Za podatke od 1960. do 2021. godine, model pokazuje najbolje performanse sa sljedećim varijablama:

- Godina
- Očekivano trajanje života
- Omjer ovisnosti o dobi, stari
- Preživljavanje do 65. godine, muškarci

Model ima sljedeće performanse na neviđenim podacima:

- MAE: 909.22
- MSE: 2202856.94
- RMSE: 1484.2
- MAPE: 0.0166

MAPE od 1.66% te MAE oko 900 ukazuju na odlične rezultate modela, što možemo vidjeti i na donjem grafu. Model ima veća odstupanja na podacima za treniranje, ali na novim podacima ima vrlo točna predviđanja te prati trend i za 2021. godinu.

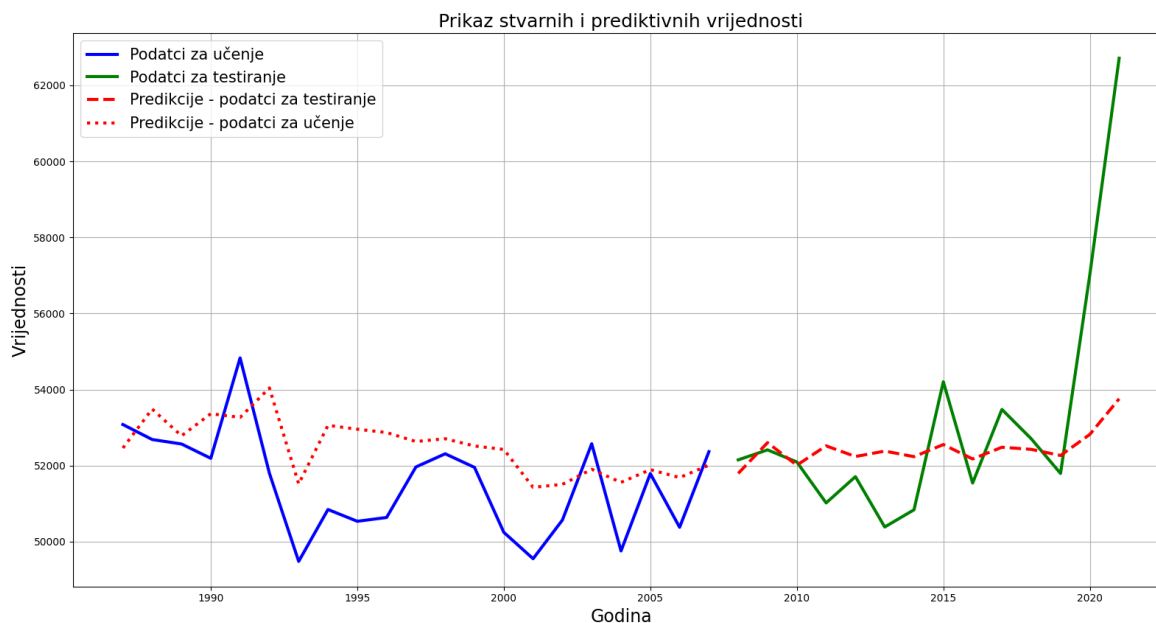


Najveći utjecaj na model ima varijabla “Preživljavanje do 65. godine, muškarci”. Uklanjanjem varijable "godina" ili varijable “Preživljavanje do 65. godine, muškarci”, performanse modela se pogoršavaju, ali još uvijek su zadovoljavajuće i prate trend. Kod ovog modela, AR i MA parametri imaju vrlo niske p-vrijednosti, što ukazuje na statistički značajne efekte tih dijelova modela.

Model sa najboljim performansama za podatke od 1986. do 2021. godine ne koristi varijablu “Preživljavanje do 65. godine, muškarci”, ali dodane su nove varijable “CPI, sezonski” te “CPI”. Model ima sljedeće performanse:

- MAE: 1659.94
- MSE: 7921903.93
- RMSE: 2814.59
- MAPE: 0.0294

Ovaj model također pokazuje vrlo zadovoljavajuće performanse, prati trend te značajnije odstupanje ima samo za 2021. godinu, što se može vidjeti i na donjem grafu.



Model najviše ovisi o varijablama “Očekivano trajanje života“, “CPI“, “CPI, sezonski“ te ima malu značajnost AR i MA parametara.

3.6. XGBoost

Za izgradnju modela korišten je 'XGBRegressor' iz biblioteke 'xgboost', koji se koristi za regresijske probleme primjenom XGBoost algoritma. Za analizu utjecaja pojedinih varijabli u modelu korištena je funkcija 'plot_importance' iz iste biblioteke.

3.6.1. XGBoost za broj rođenih

XGBoost model za broj rođenih za podatke od 1960. do 2021. godine izgrađen je sa sljedećim varijablama:

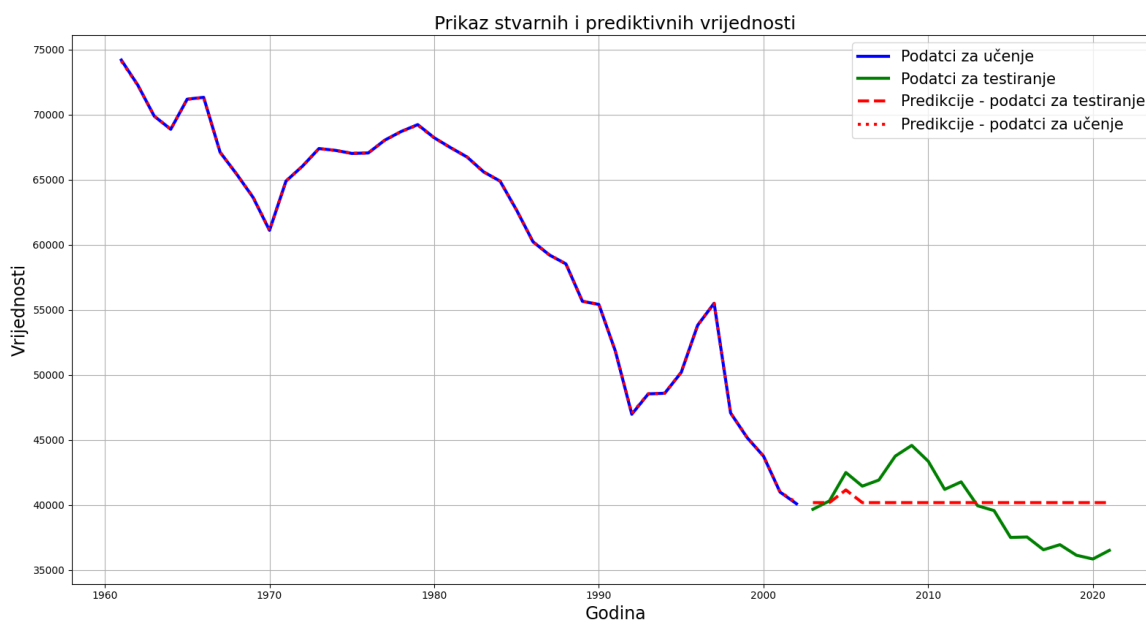
- Ruralno stanovništvo kao % ukupnog stanovništva
- Stanovništvo (15-64)
- Žensko stanovništvo (20-24)
- Godišnji rast stanovništva

Te ima sljedeće performanse na neviđenim podacima:

- MAE: 2306.17
- MSE: 7329064.94

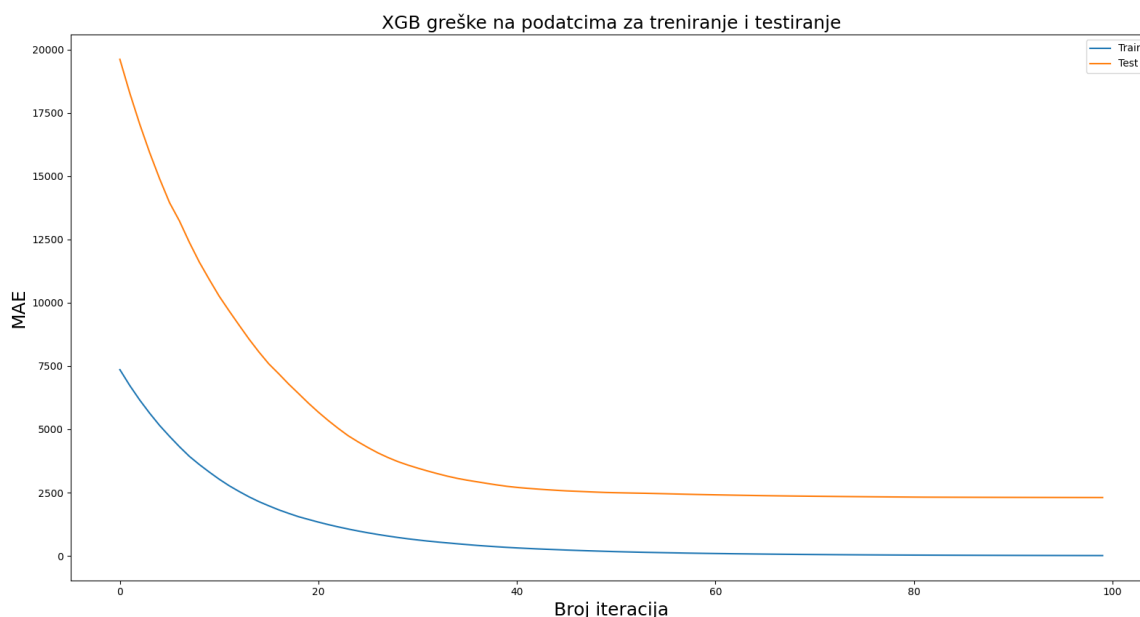
- RMSE: 2707.22
- MAPE: 0.0589

Iako su ovakvi rezultati zadovoljavajući, na sljedećem grafu možemo vidjeti da model ne prati trend.



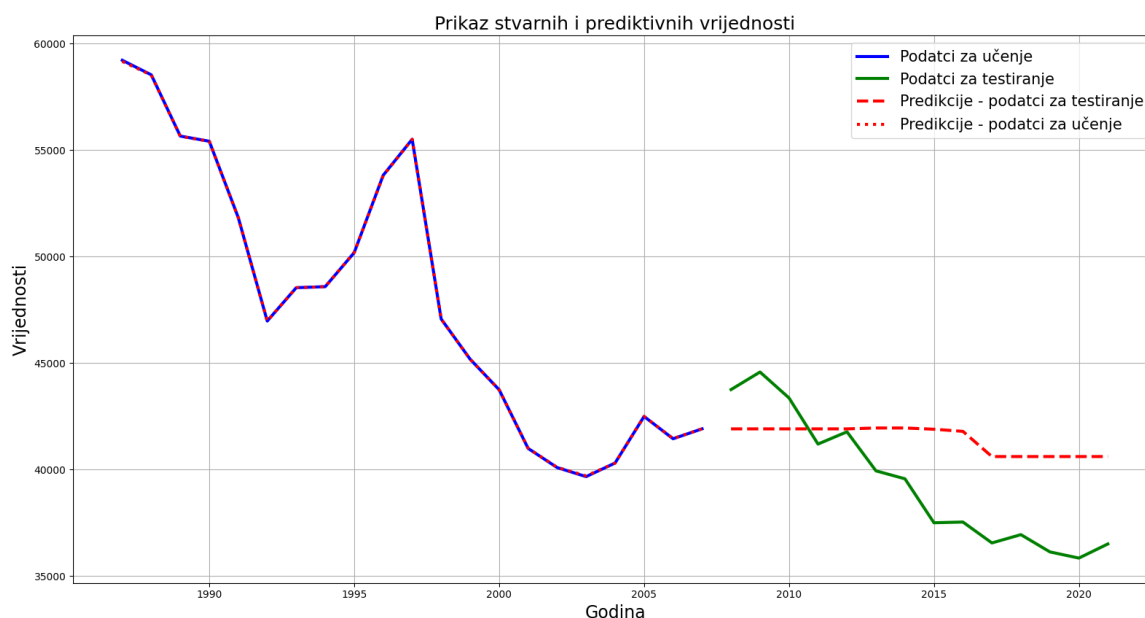
Model prati trend samo za 2005. godinu te ima konstantno predviđanje za preostale godine, što nije u skladu sa stvarnim vrijednostima koje imaju oscilacije. To ukazuje na nedostatke modela. Mogući razlozi su prenaučenost ili nedovoljna složenost.

Na sljedećem grafu možemo vidjeti greške na podacima za treniranje te podacima za testiranje. Budući da se graf pogrešaka na podacima za testiranje smanjuje te na kraju ostaje stabiliziran te nema naknadnih povećanja, možemo zaključiti da prenaučenost nije problem ovog modela.



Dodavanjem nekoliko drugih varijabli vrijednosti predikcije više nisu konstantne, ali ne prate trend te se MSE i MAPE povećavaju. Budući da varijable koje imamo za ovaj vremenski period nisu zadovoljavajuće možemo zaključiti da nam je potrebno više podataka.

Model na podacima od 1986. do 2021. ima sličan problem. Na donjem grafu možemo vidjeti njegove predikcije.



Taj model ima dvije dodatne varijable, “Urbano stanovništvo, apsolutno“ i “Godišnja promjena inflacije“.

Oba modela najviše ovise o varijabli “Ruralno stanovništvo, % ukupnog stanovništva“.

3.6.2. XGBoost za broj umrlih

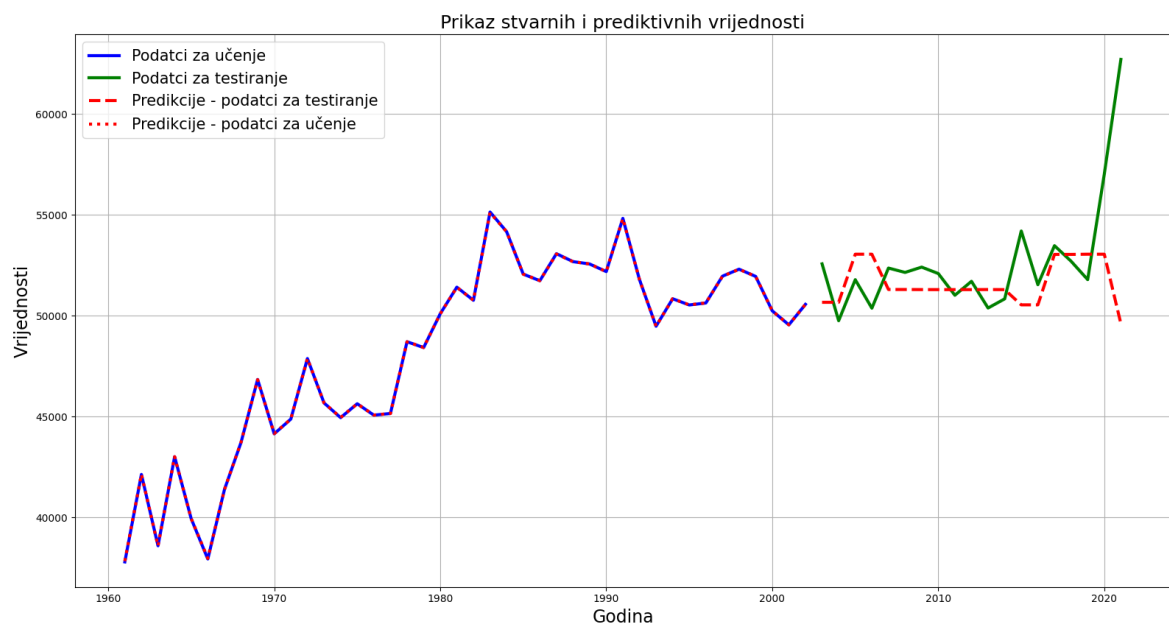
XGBoost model za broj umrlih za podatke od 1960. do 2021. godine izgrađen je sa sljedećim varijablama:

- Očekivano trajanje života
- Stanovništvo iznad 80.-te godine
- Neto migracije
- Godišnji rast stanovništva
- Stanovništvo najvećeg grada (% urbanog stanovništva)
- Omjer ovisnosti o dobi, stari

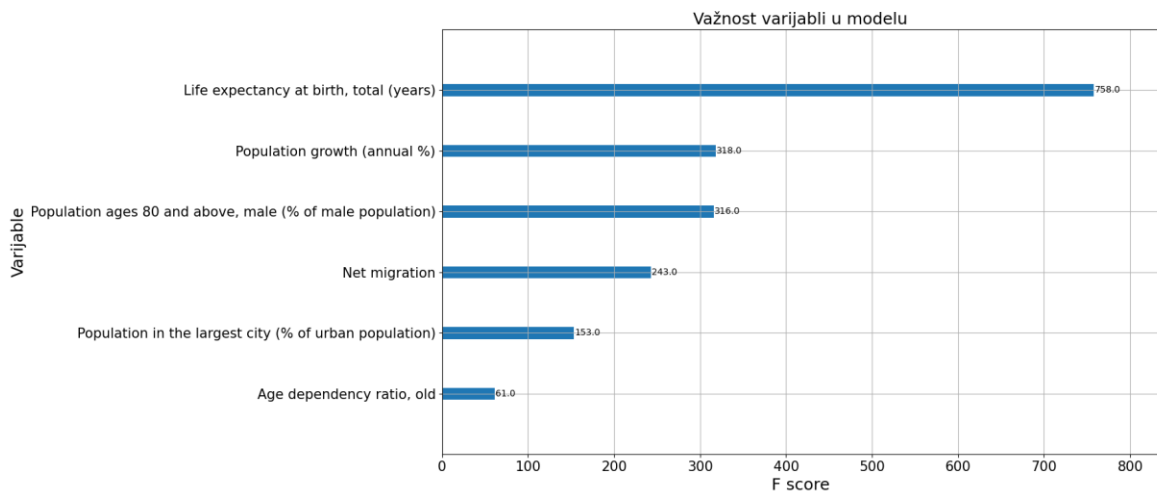
Te ima sljedeće performanse na neviđenim podacima:

- MAE: 1916.53
- MSE: 11686003.33
- RMSE: 3418.48
- MAPE: 0.0342

Ovakvi rezultati ukazuju na zadovoljavajući model što potvrđuje i donji graf. Model prati trend osim za 2021. godinu kada ima značajno odstupanje.

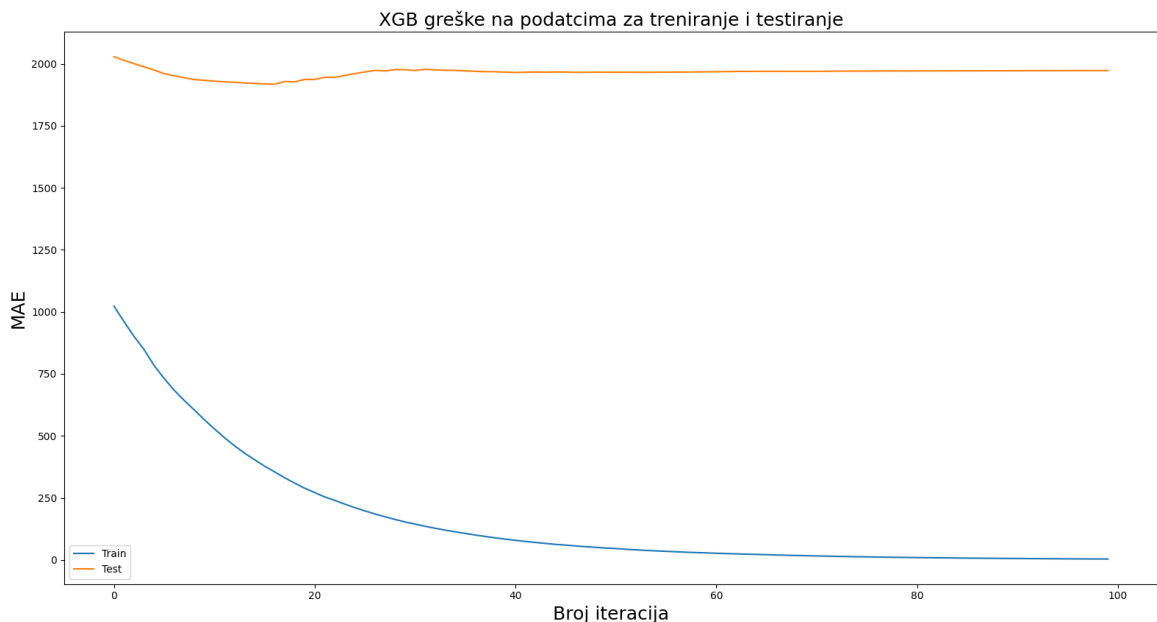


Model najviše ovisi o varijabli “Očekivano trajanje života“, dok varijabla “Omjer ovisnosti o dobi, stari“ ima mali utjecaj na model, što možemo vidjeti na donjem grafu.



Varijabla 'Preživljavanje do 65. godine, muškarci', koja je imala najveći utjecaj u modelima linearne regresije i ARIMAX-a, izostavljena je iz ovog modela jer njezino uključivanje negativno utječe na performanse modela.

Model za razdoblje od 1986. do 2021. godine s bilo kojom kombinacijom varijabli je pokazivao prenaučenosť, prikazano na donjem grafu.



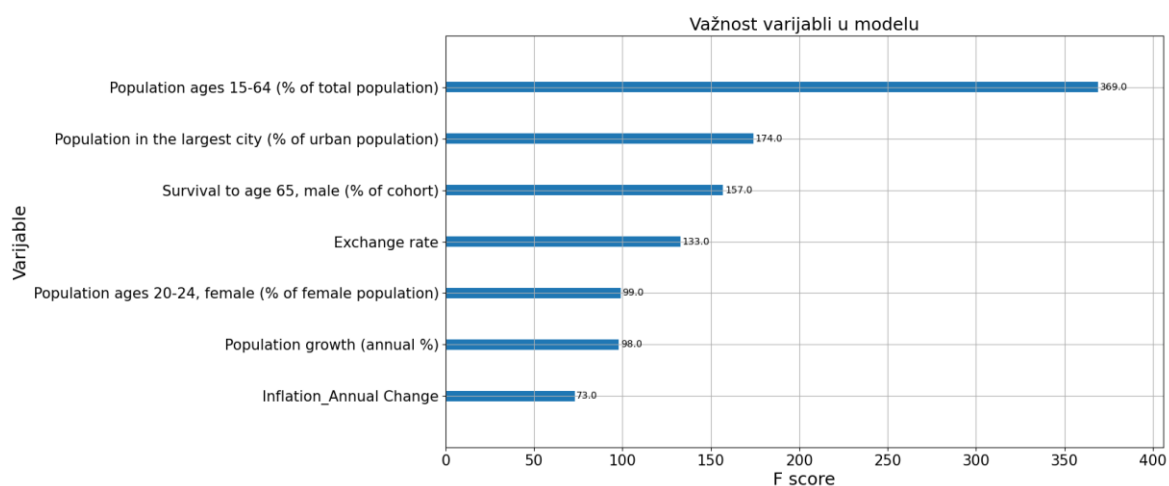
Usprkos tome pokazuje zadovoljavajuće performanse i predikcije.

Model sa najboljim performansama je izgrađen sa sljedećim varijablama:

- Stanovništvo (15-64)
- Žensko stanovništvo (20-24)

- Preživljavanje do 65. godine, muškarci
- Stanovništvo najvećeg grada kao % urbanog stanovništva
- Godišnji rast stanovništva
- Tečaj valute
- Godišnja promjena inflacije

Iako je kod prijašnjeg modela varijabla "Očekivano trajanje života" imala najveći utjecaj, u ovom modelu ona ne pokazuje nikakav utjecaj. Umjesto toga, u ovom modelu najveći utjecaj ima varijabla "Stanovništvo (15-64)", dok među ekonomskim varijablama najznačajnija je "Tečaj valute", što možemo vidjeti na donjem grafu.



4. Rezultati i diskusija

4.1. Koreliranost

Iako su brojne koreliranosti očekivane, neki od odnosa varijabli nisu posve jasni i predvidljivi na prvu te bi se kod matematičkih modela te varijable drugačije upotrebljavale za izračun predikcija. Republika Hrvatska je relativno nova država, sa različitim stupnjevima razvoja u različitim područjima što doprinosi koreliranosti koje nisu prisutne u drugim zemljama. U nastavku analizirat ćemo kako te koreliranosti utječu na razvoj prediktivnih modela.

4.2. Rezultati modela

4.2.1. Modeli za natalitet

U donjoj tablici prikazane su performanse modela na različitim skupovima podataka za predikcije varijabli nataliteta. Iz tablice možemo uočiti da najbolje performanse imaju Linearna regresija na manjem skupu podataka te Holt na cijelom skupu podataka. Budući da su ARIMAX i XGBoost složeniji modeli te generalno zahtijevaju više podataka, to može ukazivati na linearnost između varijabli i na moguće premale skupove podataka. Također zanimljivo je da modeli linearne regresije te ARIMAX-a imaju bolje performanse na manjem skupu podataka. Kod linearne to možemo pripisati dodavanju nove ekonomske varijable „CPI“, dok kod ARIMAX modela bez dodavanja novih varijabli imamo bolje rezultate. Razlog tome mogu biti fluktuacije ili lošija dokumentaciju starijih podataka. Budući da je Hrvatska mlada zemlja, moguće je i da se demografske varijable drugačije mijenjaju u novije vrijeme.

Model	Ciljna varijabla	Skup podataka	Najznačajnije varijable	MSE	MAPE
Linearna regresija	Broj rođenih	1960.-2021.	godina, stan. najvećeg grada, omjer ovisnosti dobi(mladi)	2219.53	0.056

Linearna regresija	Broj rođenih	1986.-2021.	omjer ovisnosti dobi(mladi), CPI, godina	1492.09	0.0395
Holt	Stope rođenih	1960.-2021.	-	0.43	0.0415
Holt-Winters	Stope rođenih	1960.-2021.	-	0.58	0.0596
ARIMAX	Broj rođenih	1960.-2021.	godina, stan. najvećeg grada, rast ruralnog stan.	2983.47	0.0719
ARIMAX	Broj rođenih	1986.-2021.	stan. najvećeg grada, godina, rast ruralnog stan.	2351.5	0.0606
XGBoost	Broj rođenih	1960.-2021.	% ruralnog stan., žensko stan.(20-24), rast stan.	2306.17	0.0589
XGBoost	Broj rođenih	1986.-2021.	% ruralnog stan., god. promjena inflacije, žensko stan. (20-24)	2922.82	0.0769

4.2.2. Modeli za mortalitet

U donjoj tablici prikazane su performanse modela na različitim skupovima podataka za predikcije varijabli mortaliteta. Holt model ima malo lošije performanse te ARIMAX model na cijelom skupu podataka se izdvaja kao iznimno dobar, dok modeli imaju približno jednake performanse. Za razliku od modela za predikcije nataliteta, ovi modeli ne pokazuju velike razlike između performansi na cijelom i smanjenom skupu podataka.

Model	Ciljna varijabla	Skup podataka	Najznačajnije varijable	MSE	MAPE
Linearna regresija	Broj umrlih	1960.-2021.	godina, preživljavanje do 65.-te(muškarci), očekivano trajanje života	1628.52	0.0294
Linearna regresija	Broj umrlih	1986.-2021.	godina, preživljavanje do 65.-te(muškarci)	1692.02	0.0298

Holt	Stope umrlih	1960.-2021.	-	1.08	0.0889
Holt-Winters	Stope umrlih	1960.-2021.	-	0.46	0.0333
ARIMAX	Broj umrlih	1960.-2021.	preživljavanje do 65.-te(muškarci), godina, očekivano trajanje života	909.22	0.0166
ARIMAX	Broj umrlih	1986.-2021.	CPI, CPI(sezonski), očekivano trajanje života	1659.94	0.0294
XGBoost	Broj umrlih	1960.-2021.	očekivano trajanje života, god. rast stan., muško stan. starije od 80 godina	1916.53	0.0342
XGBoost	Broj umrlih	1986.-2021.	stan.(15-64), stan. najvećeg grada, preživljavanje do 65.-te(muškarci)	1972.61	0.0345

4.3. Diskusija

Iako se u literaturi često spominje BDP kao jedan od glavnih ekonomskih varijabli za predikcije u stanovništvu, ovdje se pokazao suprotnim. Korišten je kao varijabla u samo jednome modelu, linearnoj regresiji na cijelom skupu podataka za predikcije broja rođenih, dok je za ostale modele imao većinom negativan učinak. Analizom korelacije također smo otkrili neke neočekivane veze između varijabli. Više je mogućih razloga za takva opažanja, no važno je napomenuti da je Republika Hrvatska mlada država, sa mnogo značajnih događaja u nedavnim godinama, poput rata i krize, koji uvelike utječu na demografske promjene. Također Hrvatska je u nekim područjima brzo razvijajuća zemlja, dok u drugima zaostaje, što ponovo može dovesti do raznih fluktuacija u podacima i ne neočekivanih trendova. Uz to, Hrvatska ima relativno malo stanovništvo u odnosu na države na kojima su izrađivani modeli u literaturi, što također može otežati pronalaženje trendova. Usprkos tome, većina modela pokazuje zadovoljavajuće performanse.

Modeli za mortalitet generalno imaju bolje performanse što može biti zahvaljujući većem skupu podataka vezanih za mortalitet. No također broj rođenja kroz godine ima puno veće oscilacije što može uzrokovati teži pronalazak trenda i lošije predikcije.

Stariji skupovi podataka za druge ekonomske varijable osim BDP-a, poput prosječnih primanja i potrošnje, cijene nekretnina te stope zaposlenosti, mogle bi pomoći kod budućih predikcija, posebice za natalitet.

Zaključak

Na kraju rada piše se kratak zaključak, duljine do najviše jedne stranice.

Literatura

Popis literature dolazi na kraju rada, iza zaključka, a prije ostalih priloga.

Na naslov **Literatura** primijenite stil Heading 1, a zatim ručno maknite brojčanu oznaku (to je važno kako bi i naslov „Literatura“ ušao u sadržaj na početku rada, prije uvoda).

Pri kreiranju navoda u popisu literature koristite stil *literatura*.

Primjeri u nastavku ilustriraju navođenje raznih izvora u popisu literature: (1) knjige, (2) članka u časopisu, (3) članka u zborniku konferencije, (4) doktorskog, magistarskog ili diplomskog rada, (5) web-stranice.

- [1] Tanenbaum, A. S., Wetherall, D. J. *Computer Networks*. 5. izdanje. London: Pearson, 2013.
- [2] Brady, P.T. *A Statistical Analysis of On-off Patterns in 16 Conversations*, Bell System Technical Journal, 47,1 (1998), str. 55-62.
- [3] Brady, N. *A Statistical Analysis of Use Case*. Proceedings of the 7th International Conference on Telecommunications ConTEL, Zagreb, (2003), str. 45-52.
- [4] Ivić, M. *Analiza ponašanja korisnika u digitalnim igrima namijenjenima učenju*. Diplomski rad. Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, 2016.
- [5] Epstein M., *The best VR headset in 2019*, PC Gamer, (2019, listopad). Poveznica: <https://www.pcgamer.com/best-vr-headset/>; pristupljeno 4. listopada 2019.

Uz svaki preuzeti sadržaj u svom radu – bilo da je riječ o tekstu (izravno citiranome ili „prepričanome“), slici ili grafičkom prikazu – treba navesti oznaku izvora (članak, knjiga, web-stranica ...) u popisu literature te se na nju „pozvati“, na primjer:

Međusobno povezivanje mreža zasniva se na primjeni komunikacijskih protokola (Tanenbaum i Wetherall, 2014).

Podaci o karakteristikama uređaja za virtualnu stvarnost preuzeti su s portala PC Gamer [5].

Početna verzija programa preuzeta je iz diplomskog rada [5].

U danim primjerima mogli ste uočiti dva načina referenciranja:

- (Tanenbaum i Wetherall, 2014),
- [1].

Kad izaberete jedan od njih svakako ga se držite konzistentno u cijelome radu.

Sažetak

Naslov, sažetak, ključne riječi (na hrvatskom jeziku)

Sažetak opisuje sadržaj rada, prepričan u stotinjak riječi.

Summary

Title, summary, keywords (na engleskom jeziku)