* Pozdrav
* Tema, cilj, očekivanja, svrha
* Pregled literature? (ili unutar preksljedeće natuknice)
* Podatci, analiza podataka
* Koji su moguci nacini/modeli za ovakav problem
* Koje sam modele odabrala i zašto
* Kaj sam uspjela napravit, kaj ne, koji model kak kaj zake
* Moguci razlozi za rezultate kakvi vec jesu
* Budućnost

Bok dej! Valentina desu ka!

Tema mog rada je prediktivno modeliranje nataliteta i mortaliteta u RH na temelju demografskih i ekonomskih varijabli. Dobiveni prediktivni modeli koriste se za razumijevanje utjecaja pojedinih značajki na stope nataliteta odnosno mortaliteta te se mogu koristit i za predviđanje budućih vrijednosti/trendova.

Rad je posebno koristan za donositelje odluka unutar populacijskih politika te kao pomoć pri izgradnji obrazovnog, zdravstvenog i mirovinskog sustava. Razumijevanje ovisnosti stopa nataliteta odnosno mortaliteta te drugih demografskih i ekonomskih značajki može pomoći u planiranju i prilagodbi tih sustava.

Očekivanja od rada su izgradnja točnih prediktivnih modela te uvidi u pozitivne i negativne utjecaje značajki na stope nataliteta i mortaliteta.

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 jeve matrice i Gompertz ovog modela. No takvi modeli mogu biti ograničeni pri velikim skupovima podataka te pronalasku nelinearnih odnosa između značajki.

Posljednjih godina za ovakve probleme se sve više koriste modeli strojnog učenja koji mogu prepoznati različite obrasce i odnose između značajki što omogućuje preciznije predikcije. Neki od najčešće korištenih su linearna regresija, Bayesov model, ARIMA, stabla odluke, Gradient Boosting modeli te neuronske mreže. Najčešće korištena značajka u drugim radovima uz same varijable nataliteta i mortaliteta je BDP.

--- prednosti i nedostatci tih metoda

U ovom radu su odabrani modeli linearne regresije, Holt, Holt-Winters, ARIMAX te Extreme Gradient Boosting.

Linearna regresija zbog svoje jednostavnosti te identifikaciju linearnih odnosa između varijabli, Holt zbog sposobnosti da uoči trendove , ARIMAX i XGBoost zbog mogućnosti uočavanja složenijih nelinearnih odnosa.

Prije izrade samih modela, napravila sam analizu korelacije između broja rođenih odnosno umrlih te drugih demografskih i ekonomskih značajki.

Ovaj graf predstavlja korelacije između broja rođenih i drugih značajki

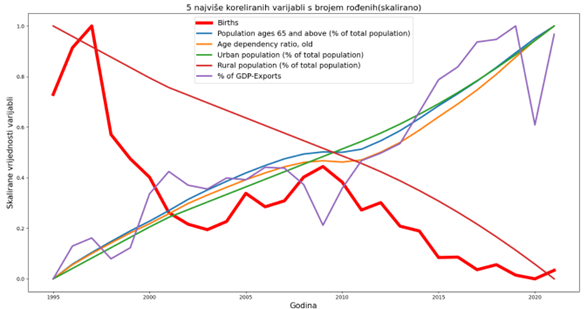
A chart with different colored lines

Description automatically generated with medium confidence

Imamo neke očekivane odnose poput negativne korelacije sa starom populacijom, indeksom potrošačkih cijena, urbanim stanovništvom, pozitivnu korelaciju sa mlađim i ruralnim stanovništvom.

Zanimljiviji odnosi su negativna korelacija s BDP-om te postotkom izvoza, to se može objasniti kroz mogući veći razvoj države, urbanizaciju, veća uključenost žena na tržištu rada, više obrazovanje, veću dostupnost kontracepcije što može negativno utjecati na manji broj rođenih, dok će BDP odnosno izvoz rasti

Na ovom grafu možemo vidjeti kako se vrijednosti 5 najviše koreliranih varijabli i broj rođenih mijenja kroz vrijeme. Odnosno vidimo da pozitivno korelirana varijabla ruralnog stanovništva više-manje prati trend broja rođenih, dok ostale varijable s negativnom korelacijom zapravo prate obrnuto proporcionalni trend.



Druga analiza je znači analiza korelacija značajki sa brojem umrlih, rezultate možemo vidjeti na ovom grafu:

A chart with different colored lines

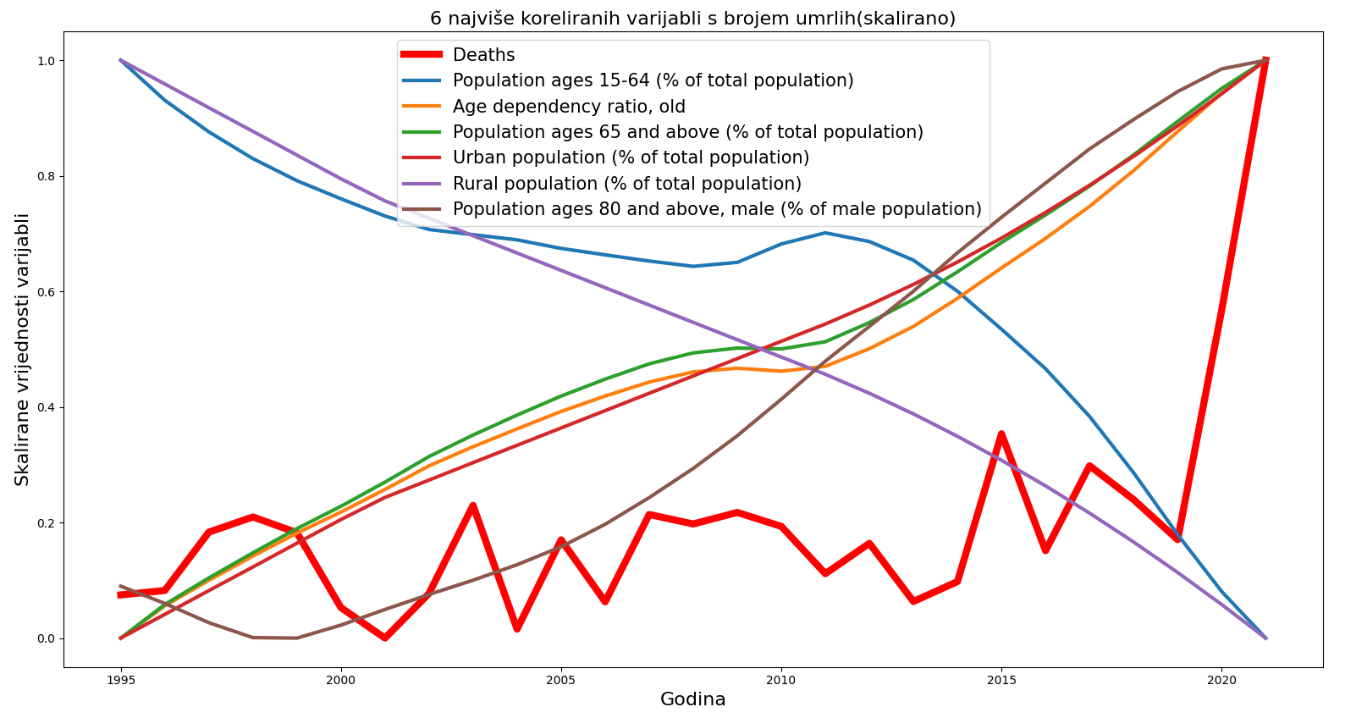
Description automatically generated with medium confidence

Ponovo možemo vidjeti očekivane odnose, pozitivnu korelaciju sa starijim stanovništvom, negativnu korelaciju sa mlađim stanovništvom

Od zanimljivijih odnosa imamo urbanu populaciju s pozitivnom korelacijom, ruralnu s negativnom korelacijom, što iako veći udio urbane populacije odnosno manji ruralne 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

BDP i izvoz pozitivno korelirani s brojem umrlih, Veći BDP i izvoz 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 varijabli kroz vrijeme.



Ponovno vidimo da varijable s negativnom korelacijom su obrnuto proporcionalne, a one s pozitivnom proporcionalne broju umrlih.

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

Prvi model je model linearne regresije je za broj rođenih na razdoblju od 1960. do 2021. 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

Te na ovom grafu možemo vidjeti da model nije savršen, uspio uhvatiti opći trend, te postoje odstupanja u predviđanjima, posebice za 2021 godinu.

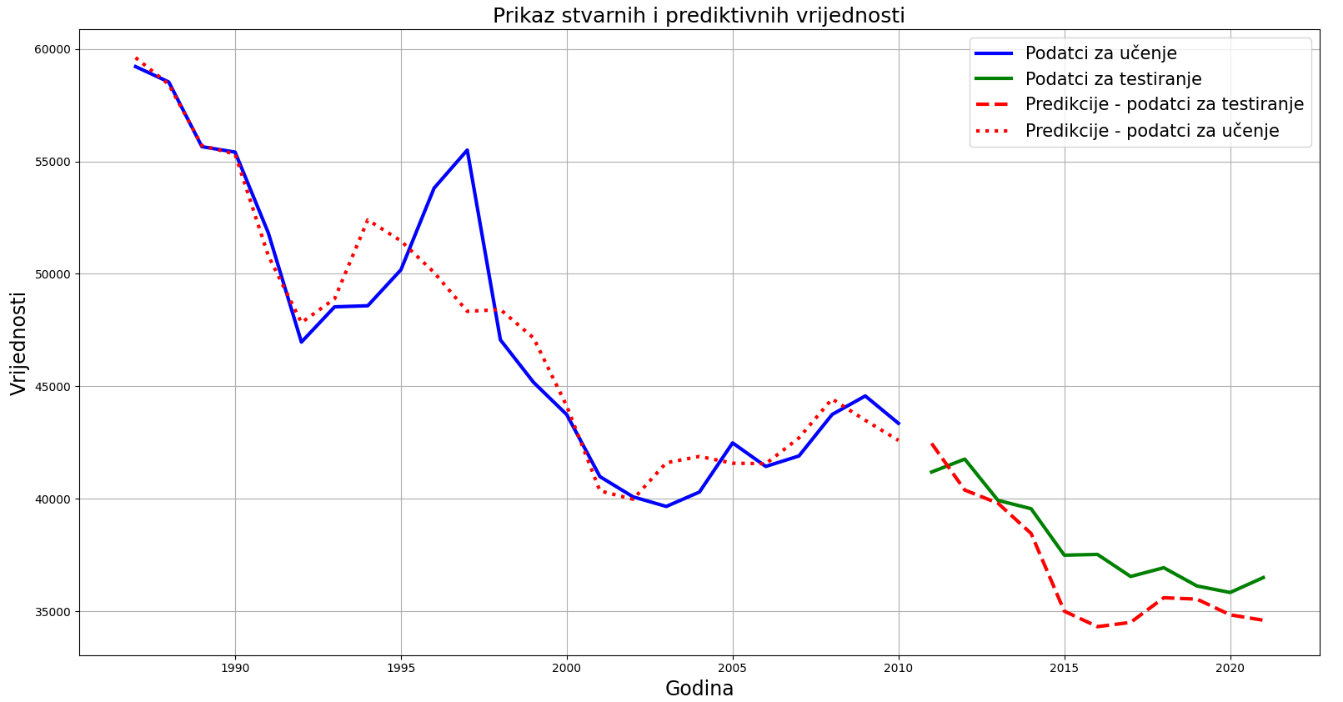
A graph with lines and numbers

Description automatically generated

Za model na manjem skupu podataka od 1986. do 2021. godine dodane su varijable:

* CPI, sezonski (indeks potrošačkih cijena)
* Tečaj valute

Model ima malo bolju prosječnu apsolutnu postotnu pogrešku i apsolutnu prosječnu pogrešku.

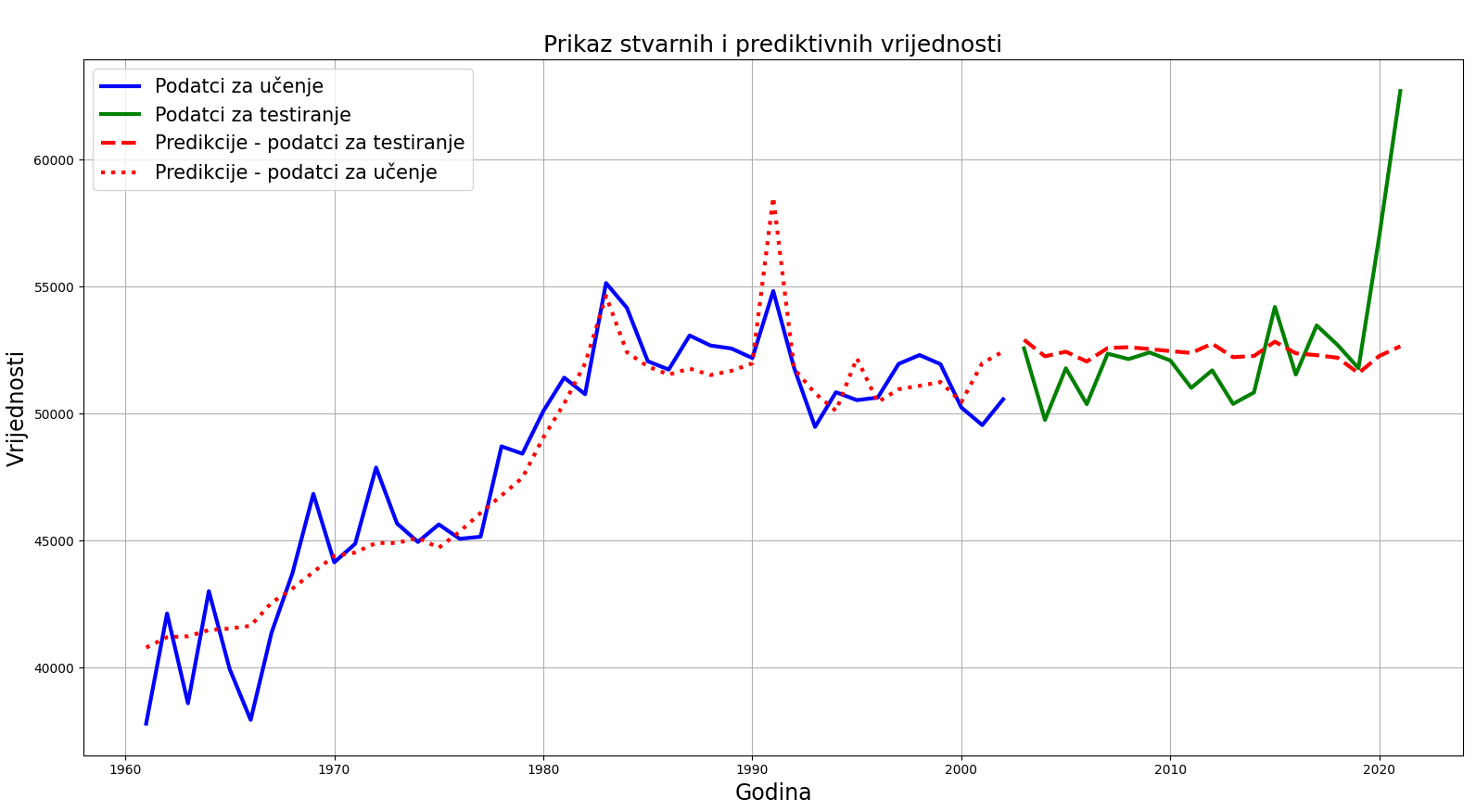
Na grafu možemo vidjeti da model malo bolje od prijašnjeg prati trend te nema toliko odstupanje u 2021. godini  


Modeli linearne regresije za broj umrlih su postigli dosta zadovoljavajuće performanse, na oba skupa podataka sa MAPE ispod 3%

Na podatcima od 1960. korištene su sljedeće varijable:

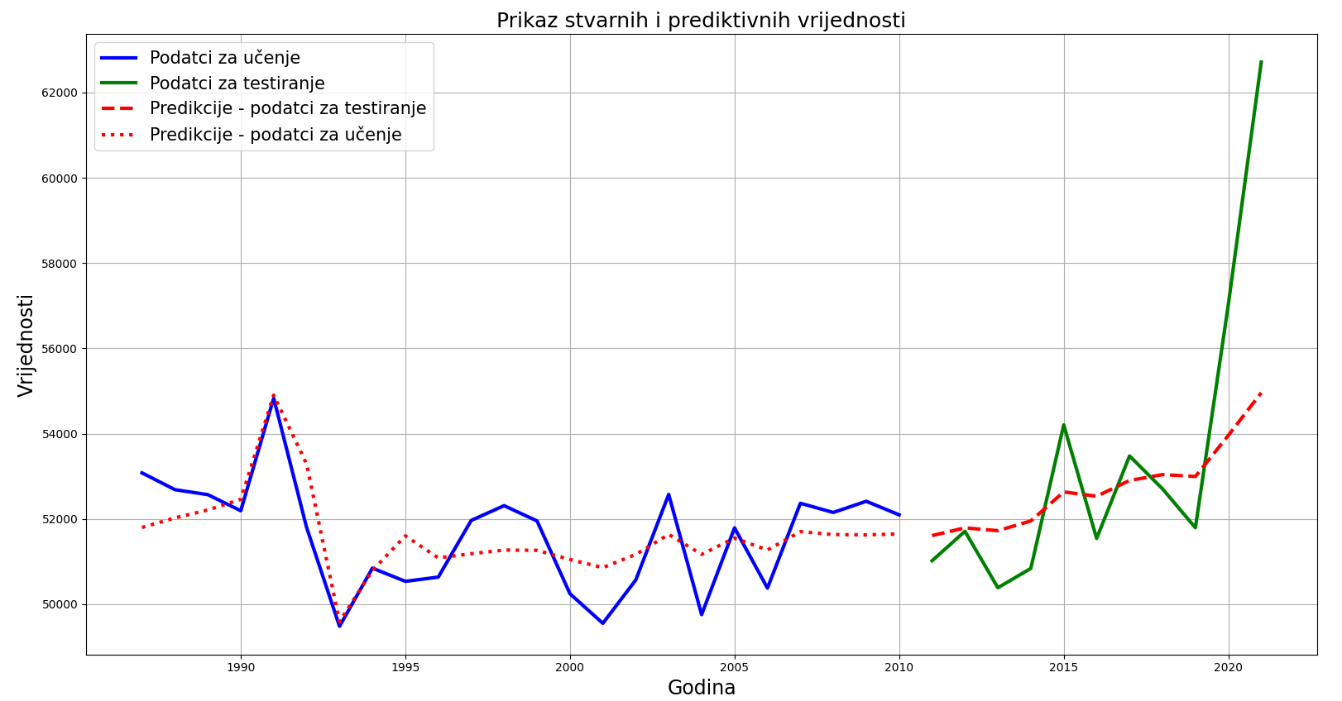
* Godina
* Omjer ovisnosti o dobi, stari
* Očekivano trajanje života
* Preživljavanje do 65. godine, muškarci

Na ovom grafu vidimo da model prati trend te ima odstupanje samo za 2020 i 2021 godinu.



Model na periodu od 1986. godine ima dodatne varijable:

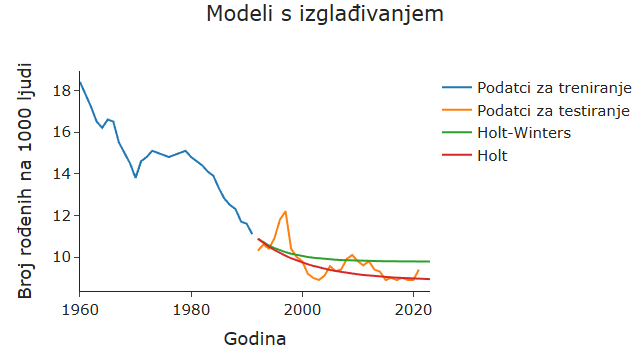
* CPI
* CPI, sezonski
* Te uklonjenu Age dependency varijablu

Vidimo da ovaj model ne prati trend malo lošije te također ima odstupanje za 2020 i 2021. godinu iako manje. 

**HOLT**

Holt i Holt-Winters modeli koriste samo prijašnje podatke ciljne varijable kao ulaz.

Ovo je graf modela za stope rođenih



Možemo primjetiti da Holtov model prati trend bolje, no niti jedan od modela nema značajna odstupanja. Kod oba modela korišten je parametar 'damped\_mod = True' što predstavlja prigušenje, odnosno 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.

Model za stope umrlih

Možemo vidjeti da Holt-Winters ima bolje performanse te prati ukupni trend s manjim odstupanjima.

A graph with different colored lines

Description automatically generated

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.

-----------------------------------------------------------------------------------------------------------------------------

**ARIMAX**

60% podataka za treniranje te preostalih 40% za testiranje.

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

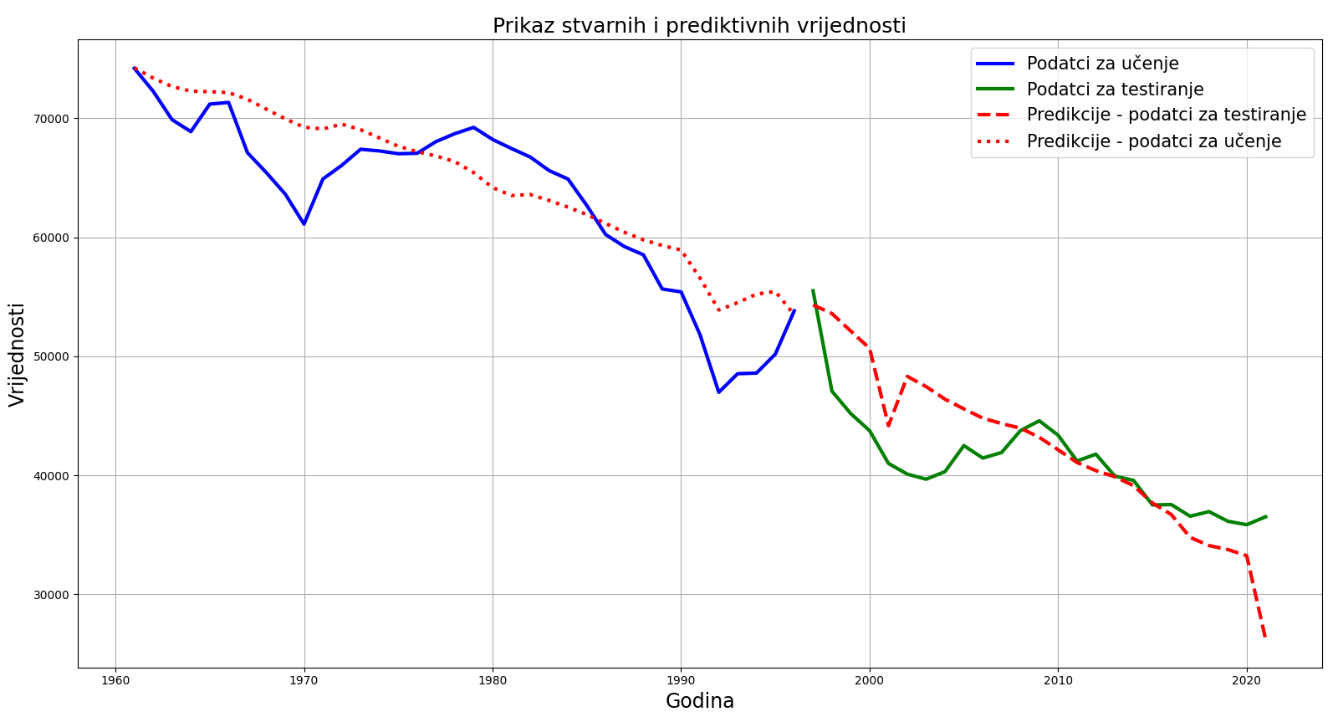
• Godina

• Neto migracije

• Godišnji rast ruralnog stanovništva

• Stanovništvo najvećeg grada

Model uočava trend, ali ne savršeno što možemo vidjeti i na sljedećem grafu. Također uočavamo da model ima najveće odstupanje za 2021. godinu.



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 podatcima od 1960. do 1986. koje su dovele do smanjenih performansi iako je skup podataka veći. Predviđanja tog modela možemo vidjeti na sljedećem grafu

A graph with lines and numbers

Description automatically generated with medium confidence

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 podatcima za učenje. Kao i kod prijašnjeg modela, utjecati AR i MA parametara su zanemarivi.

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

Ovaj model od svih ima najbolje performanse, uz MAPE: 0.0166, MAE: 909.22.

A graph with lines and a red line

Description automatically generated with medium confidence

Model ima veća odstupanja na podatcima za treniranje, ali na novim podatcima ima vrlo točna predviđanja te prati trend i za 2021. godinu. 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“.

A graph with lines and a red line

Description automatically generated with medium confidence

Ovaj model također pokazuje vrlo zadovoljavajuće performanse, prati trend te značajnije odstupanje ima samo za 2021. godinu, MAPE: 0.0294 MAE: 1659.94

Ima malu značajnost AR i MA parametara.

**XGBOOST – mzd koje sam parametre tu koristila?**

1960. do 2021. godine ostvarene su uz 70% podataka za treniranje te 30% za testiranje, a za period od 1986. do 2021. godine uz 60% podataka za treniranje odnosno 40% za testiranje.

Broj rođenih

varijablama:

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

perf

* MAPE: 0.0589
* MAE: 2306.17

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

A graph with lines and dots

Description automatically generated with medium confidence

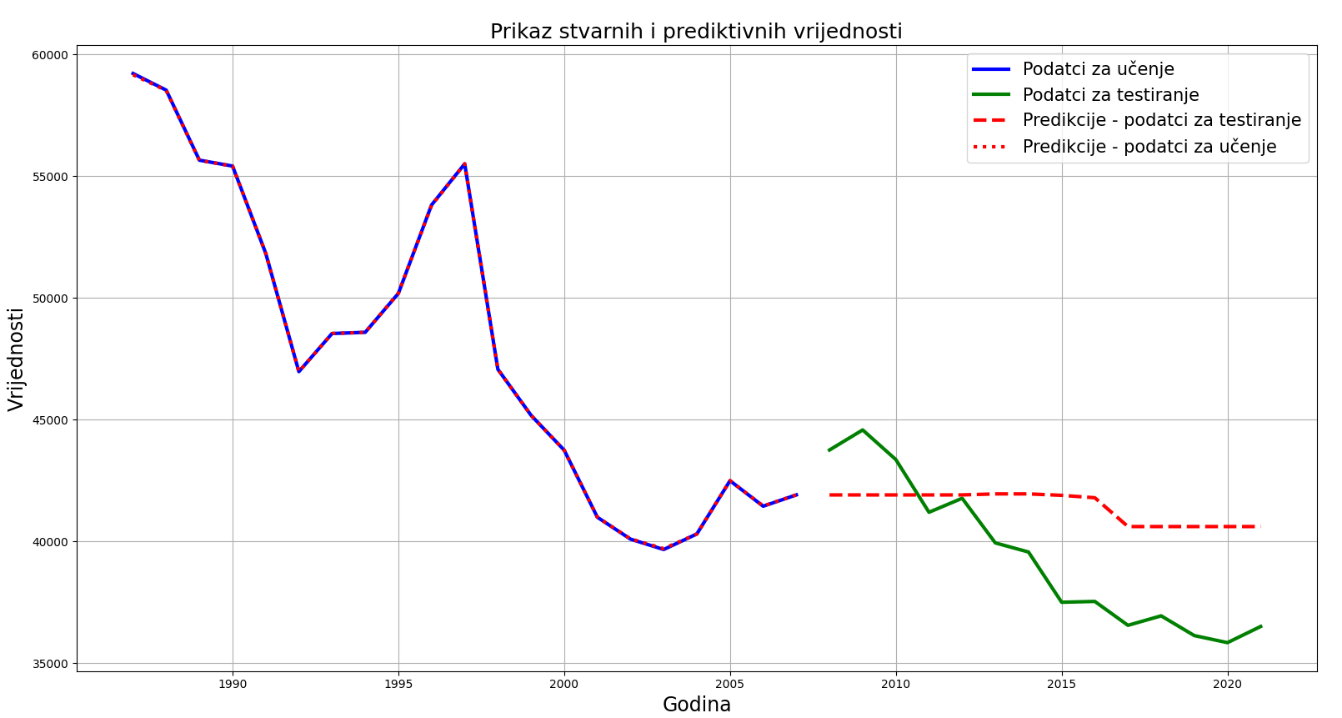
Model prati trend samo za 2005. godinu te ima konstantno predviđanje za preostale godine, što nije u skladu sa stvarnim vrijednostima koje imaju oscilacije. To ukazuje na nedostatke modela. Mogući razlozi su prenaučenost ili nedovoljna složenost.

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

A graph of a person with a line

Description automatically generated with medium confidence

Model na podatcima od 1986. do 2021. ima sličan problem.



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

BROJ UMRLIH

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

• MAPE: 0.0342

• MAE: 1916.53

Ovakvi rezultati ukazuju na zadovoljavajući model što potvrđuje i donji graf (Slika 3.21). Model prati trend osim za 2021. godinu kada ima značajno odstupanje.

A graph with lines and numbers

Description automatically generated with medium confidence

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

A graph of a person with a blue line

Description automatically generated with medium confidence

ZAKE NEMAM GRAFA ZA TAJ MODEL ??????????????????

USPOREDBA ITD.

A close-up of a box

Description automatically generated

A table with numbers and text

Description automatically generated

najbolje performanse imaju linearna regresija na manjem skupu podataka te Holt na cijelom skupu podataka.

Najčešće varijable najbitnije kaj već: stan. najvećeg grada, ruralno stanovništvo,

### Modeli za mortalitet

A screenshot of a table

Description automatically generated

FALI MI U ZAVRSNOM U RECENICI „OSTALI“,

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

**ZAKLJUCAK ZAVRSETAK NEKE NEPFAOKLFWLAČWAĆ**

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 podatcima i 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**

Rad je pokazao da su manje složeni modeli poput linearne regresije i Holt modela zadovoljavajući za ovakve probleme te da složeniji modeli poput ARIMAX-a i XGBoost-a više ovise o veličini dostupnih skupova podataka.

Modeli za mortalitet su općenito pokazali bolje performanse zbog većeg broja povezanih varijabli te manjih oscilacija u samim podatcima broja umrlih u odnosnu na broj rođenih. Također analiza podataka otkrila je neke neočekivane korelacije, najistaknutije one između BDP-a i broja rođenih, odnosno broja umrlih kao mogući rezultat specifičnih povijesnih i socioekonomskih uvjeta u Republici Hrvatskoj.

Iako su modeli zadovoljavajući, postoji potencijal za bolja predviđanja uključivanjem većih skupova podataka sa dodatnim varijablama, posebice za predviđanje nataliteta. Također u budućnosti, ako dođe do određene stabilizacije u ekonomskim i demografskim uvjetima, predviđanja će biti lakša i preciznija.