# Uticaj globalnog zagrevanja na porast nivoa mora i njegova predikcija

Stefan Beljić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
stefan.beljic@uns.ac.rs

Stefan Savić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
stefan.savic98@uns.ac.rs

Stefan Arađanin
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
stefan.aradjanin@uns.ac.rs

***Apstrakt* - Poslednjih godina svedoci smo sve većeg uticaja globalnog zagrevanja na našu planetu. Ono predstavlja povećanje prosečne temperature Zemljine atmosfere i okena, čime je uslovljen i porast nivoa mora na koji utiču mnogobrojni faktori. Porast nivoa mora je globalno ekološko pitanje i pretnja po obalne zajednice koje se obično karakterišu velikom gustinom stanovništva, kao i infrastrukture. Posledice globalnog zagrevanja mogle bi biti katastrofalne. Stoga, procena porasta nivoa mora, kao i detaljnija analiza faktora kojima su ove promene uslovljene, od velike su važnosti za održiv razvoj priobalnih zajednica i ostatak čovečanstva. U ovom radu, prezentovan je uticaj različitih činilaca na mesečni porast nivoa mora, kao i njegova fluktuacija tokom vremena. Skup podataka čine faktori koji su značajni za globalno zagrevanje, a to su između ostalog temperatura, otapanje glečera, gustina mora, salinitet i drugo, a koji su međusobono povezani zajedničkim datumom merenja. Takođe, u prvom delu rada obučeni su sledeći modeli: SVM, *Naive-Bayes*, *Random Forest*, *XGBoost* i *Bagging*, dok su skupovi podataka podeljeni na obučavajući i testni skup. *XGBoost* se ispostavio kao najuspešniji model, zahvaljujući njegovoj optimiziaciji hiper-parametara. Drugi deo rada bazirao se na rešavanju *Time Series* problema, koji se pokazao kao odličan pristup za ostvarenje željene predikcije nivoa mora, uz minimalan procenat greške. Skup podataka je u ovom slučaju podeljen na obučavajući, validacioni i testni skup. Kao što se i očekivalo, glavni uticaj na porast nivoa mora predstavlju emisija štetnih gasova tj. ugljen-dioksida i temperature, kako na površini vode, tako i na kopnu.**

***Ključne reči - GMSL; globalno zagrevanje; nivo mora; Time Series; model; temperatura; predikcija; XGBoost***

## I. Uvod

Globalno zagrevanje predstavlja ozbiljan problem i veliki izazov pred kojim se našlo čovečanstvo u 21. veku. Ono predstavlja konstantno povećanje prosečne temperature na Zemlji, prouzrukovano najčešće, ali ne i nužno, ljudskim faktorom. Među glavne uzroke spada povećano emitovanje štetnih gasova i ugljen-dioksida [5]. Jedna od posledica globalnog zagrevanja jeste učestalo povećanje nivoa mora i okeana. Na ovaj rast utiču mnogobrojni faktori kao što su: temperatura kopna i mora, otapanje glečera, površinska gustina mora i drugo. Promena obale, sve veća učestalost i intenzitet poplava i podzemnih voda su samo neki od fizičkih uticaja ovog konstantnog povećanja, dok društveno-ekonomske posledice uključuju degradaciju kvaliteta vode, oštećenje infrastrukture, iscrpljenje poljoprivrednih resursa i drugo.

U ovom radu će biti predstavljena dva rešenja za određivanje porasta nivoa mora. Jedno rešenje predstavlja odgovor na pitanje da li se desilo povećanje nivoa mora u odnosu na prethodni mesec i koji faktori na njega najviše utiču. Ovaj pristup je realizovan kroz različite modele obučene na osnovu najčešćih faktora globalnog zagrevanja. Drugo rešenje koje je predstavljeno u ovom radu je *Time Series* problem, koji predviđa tačnu promenu srednjeg globalnog nivoa mora (GMSL). Oba načina predstavljaju dobru osnovu za dalje rešavanje ovog ozbiljnog problema globalnog zagrevanja.

Mnogobrojni izazovi javili su se prilikom rešavanja ovog problema. Prikupljanje podataka predstavlja najveću prepreku. Verodostojni i precizno iskazani podaci veoma su zaštićeni i mereni od strane visoko nadležnih institucija i organizacija. Neusklađenost, nekonzistentnost i neredovno merenje, dovelo je do velikog broja nedostajućih vrednosti prilikom njihovog povezivanja. Iz ovog razloga, većina radova se fokusira na rešavanje lokalnog, a ne globalnog problema porasta nivoa mora. U radu je dokazano i da se sa oshudnim brojem podataka, mogu dobiti zadovoljavajuće precizni rezultati, koji bi trebalo da daju podstrek za masovnija istraživanja.

U narednom poglavlju, opisana su srodna istraživanja i naučni radovi. Zatim, predstavljeni su opis i analiza skupa podataka. U četvrtom poglavlju opisane su metodologije koje su korišćene pri rešavanju problema, dok su u petom poglavlju prikazani dobijeni rezultati. Na kraju, izveden je zaključak o samom radu i problemu globalnog zagrevanja.

## II. Srodna istraživanja

U radu [1], Balogun i Adebisi vrše predviđanje nivoa mora pomoću neuronskih mreža kako bi procenili uticaj ansambla okeansko-atmosferskih procesa na tačnost modela. Ova studija integriše širok spektar okeansko-atmosferskih varijabli

za predviđanje varijacija nivoa mora. Poređenje modela obučenih na okeanskim sa modelima obučenim na atmosferskim varijablama, pokazuje da atmosferski procesi imaju veći uticaj na predviđanje modela od okeanskih. Međutim, obuka sa kombinovanim okeansko-atmosferskim varijablama je poboljšala predviđanja modela. Atmosferski podaci koji su korišćeni uključuju padavine, oblačnost i brzinu vetra, dok podaci o moru uključuju temperaturu, salinitet i gustinu površine mora.

Zheng u radu [2] predstavlja konstrukciju pouzdanih statističkih modela zasnovanih na ogromnim klimatskim podacima, vršeći identifikaciju odnosa između temperature i potencijalnih faktora, kao što su koncentracija ugljen-dioksida ($CO_2$), azot-oksida ($N_2O$) i metana ($CH_4$). Takođe, vrši se identifikovanje faktora koji doprinose globalnom zagrevanju. S obzirom da podaci o prosečnoj temperaturi nisu međusobno usklađeni, korišćena je linearna interpolacija za njihovo usklađivanje i efikasniju obradu. Tokom procesa obuke, korišćena je unakrsna validacija za traženje odgovarajućih hiper-parametara modela. Zatim su upoređena tri različita algoritma mašinskog učenja (*Random Forest*, *Lasso* i SVR). *Random Forest* je nadmašio ostale korišćene algoritme mašinskog učenja i pokazao je da na promenu temperature najviše utiče $CO_2$. Po uzoru na ovaj rad, iskorišćen je podatak o koncentraciji ugljen-dioksida, *Random Forest* kao jedan od algoritama mašinskog učenja i unakrsna validacija primenjena prilikom traženja odgovarajućih hiper-parametara modela.

Kroz rad [3] prikazane su napredne statističke analize, uključujući metode mašinskog učenja koje mogu pružiti koristan uvid u promene nivoa mora. Glavni cilj ovog rada je potreba za određivanjem tendencija povećanja i smanjenja nivoa mora u narednim godinama uz pomoć mašinskog učenja. U radu je korišćen model regresije Gausovog procesa (GP) i rekurzivna neuronska mreža (RNN). Analiziran je uticaj otvorenog okeana, sa njegovom promenom prosečne temperature do 700m dubine, na obalu. Kroz ovaj rad je predloženo da bi skup podataka trebalo biti proširen i dodatnim parametrima, kao što je promena debljine leda, salinitet i uticaj globalnog zagrevanja na površinske temperature, što je i primenjeno u okviru ove studije.

U radu [4] vrši se razmatranje o primeni tehnika mašinskog učenja na neke probleme regresije koji se obično sreću prilikom analize vremenskih serija podataka. S obzirom da su regionalne varijacije u porastu nivoa mora velike i moraju se uzeti u obzir kada se planira budući porast nivoa mora, koriste se istorijski podaci o nivou mora sa merača plime i oseke koji se nalaze na različitim lokacijama širom švedske obale. Međutim, metoda je univerzalna i može se isto tako lako primeniti na podatke sa drugih lokacija. Metode mašinskog učenja koje su primenjene predstavljaju tri razlite veštačke neuronske mreže (ANN) i višestruku linearnu regresiju. Pokazalo se da su ukupne performanse ovih algoritama mašinskog učenja dobre, često nadmašujući one od mnogo skupljih numeričkih modela okeana.

## III. Opis skupa podataka

U ovom poglavlju biće detaljno opisan postupak formiranja konačnog skupa podataka na osnovu više manjih skupova neophodnih za analizu i dodatnu obradu podataka.

S obzirom da na globalno zagrevanje, a samim tim i na porast nivao mora, utiču različiti faktori, nakon detaljnog istraživanja odlučeno je da se u konačnom skupu podataka uzmu u obzir sledeći atributi prikupljeni iz različitih izvora:

- ukupna površina morskog leda izražena u jedinici $10^6$ $km^2$ (Extent),
- temperatura vode izražena u celzijusima (WaterTemp),
- zasićenost vode kiseonikom (O2ml),
- koncentracija silikata (SiO3),
- koncentracija nitrata (NO3),
- globalna prosečna temperatura kopna za dan merenja izražena u celzijusima (LandAverageTemperature),
- globalna prosečna temperatura kopna i okeana za dan merenja izražena u celzijusima (LandAndOceanAverageTemperature) i
- prosečna mesečna molska frakcija $CO_2$ određena iz dnevnih proseka (CO2).

Na osnovu podataka prikupljeni od 1978. do 2015. godine sa sajta „National Snow & Ice Data Center“ [14] formiran je skup podataka u CSV formatu koji sadrži sledeće atribute: *Year*, *Month*, *Day*, *Extent*, *Missing*, *Source data*, *Hemisphere*. Podaci koji su od interesa za ovaj rad su:

- *Year* - godina merenja,
- *Month* - mesec merenja,
- *Day* - dan merenja i
- *Extent* - ukupna površina morskog leda izražena u jedinici $10^6$ $km^2$.

Kako bi se dobio željeni skup podataka sa prethodno navedenim podacima, neophodna je njegova obrada. Na početku je učitan inicijalni skup podataka iz kojeg su izbačeni podaci koji nisu od interesa (*Missing* i *Source data*). Nakon toga, izvršeno je spajanje kolona koje predstavljaju godinu, mesec i dan u jednu kolonu koji predstavlja celokupan datum, dok su pojedinčane kolone sa informacijama o godini, mesecu i danu izbrisane. Nakon analize, dolazi se do informacije da ovaj skup podataka ima 26354 redova, od toga 13177 duplikata iz razloga što za isti dan postoji zasebno zabeležena debljina leda na severnoj i južnoj hemisferi, te je bilo neophodno grupisati podatke i izračunati srednju vrednost. Nakon toga, dobijen je skup podataka koji sadrži jedinstvene vrednosti. Kako bi se to dokazalo, izvršena je provera vrednosti i uočeno je da ne postoje autlajeri. Uzimajući u obzir ostale skupove podataka o kojima će biti reči u nastavku, izdvojeni su podaci od 1969. do 2013. godine. Kako inicijalni skup podataka sadrži podatke od 1978. godine, dodate su godine koje nedostaju. S obzirom da kod datuma dani nisu neophodni, podaci su grupisani po mesecima. Na kraju, izračunati su i dodati nedostajući podaci za *Extent* od 1969. do 1978. godine.

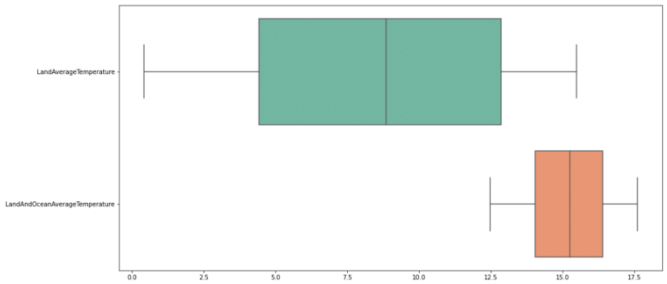
Prikupljeni su podaci o temperaturi vode (*WaterTemp*), zasićenosti vode kiseonikom (O2ml), koncentraciji silikata (SiO3) i nitrata (NO3) [10]. S obzirom da za ovaj rad nisu bitni svi podaci iz inicijalnog skup, izvršeno je uklanjanje kolona koje nisu bile od interesa u sličnim radovima i kod kojih je empirijskim putem dokazano da je standardna devijacija mala. Zbog daljeg razmatranja ostavljene su kolone *Depthm* (dubina na kojoj je izvršeno merenje) i *Zone* (zona u kojoj je izvršeno

merenje - *Photic/Disphotic*). Zatim, uzeti su u obzir samo podaci prikupljeni na dubinama do 200m (*Photic*), po ugledu na rad [3], gde je eksplicitno navedeno da velike dubine nemaju prevelik uticaj na porast niva mora. Analizom skupa podataka, izveden je zaključak da postoji samo 3080 različitih datuma, pa samim tim da postoje duplicirane vrednosti. Izdvajanjem duplikata, može se primetiti da oni predstavljaju merenja u istom danu na različitim dubinama. Grupisanjem po datumu i korišćenjem *median* metode, dobija se podatak o temperaturi vode i koncentraciji $O_2$ za određeni dan. Nakon grupisanja, izbačeni su duplikati i obrisane suvišne kolone. Kada su izvdojene sve potrebne kolone, uzeti su u obzir svi podaci od 1969. do 2013. godine i zatim dodati datumi koji nedostaju. Kako bi se dobili podaci za određeni mesec, izvršeno je grupisanje po mesecima. S obzirom da postoje nedostajuće vrednosti za neke atribute, odrađena je interpolacija kako bi se one popunile, a koja je prikazana na slici 3.

Naredno obrađivan skup podataka [11], sadrži infomracije o globalnoj prosečnoj temperaturi kopna, kao i globalnoj prosečnoj temperaturi kopna i okeana za dan merenja. U okviru ovog skupa podataka nalaze se podaci prikupljeni u periodu od 1750. do 2015. godine. Ovaj skup podataka sadrži sledeće atrbute: dt, *LandAverageTemperature*, *LandAverageTemperatureUncertainty*, *LandMaxTemperature*, *LandMaxTemperatureUncertainty*, *LandMinTemperature*, *LandMinTemperatureUncertainty*, *LandAndOceanAverageTemperature*, *LandAndOceanAverageTemperatureUncertainty*. Podaci koji su od interesa za ovaj rad su:

- dt - datum merenja,
- *LandAverageTemperature* - prosečna temperatura kopna za dan merenja i
- *LandAndOceanAverageTemperature* - prosečna temperatura kopna i mora za dan merenja.

Najpre, izbačene su kolone koje nisu od interesa i izvršeno je preimenovanje kolona koje su zadržane. Skup podataka poseduje 3192 reda i isti toliki broj jedinstvenih datuma. Iz toga je zaključujeno da nema duplikata. Datumi pre 1850. godine nisu razmatrani i samim tim je uklonjen dobar deo nedostajućih vrednosti. Na osnovu slike 1, uočeno je da ne postoje autlajeri i uzeti su u obzir datumi od 1969. do 2013. godine.



Slika 1. Grafički prikaz autlajera za temperature kopna i okeana

Za naše potrebe iskorišćen je skup podataka [12], u kom se pored datuma u kom je izvršeno merenje nalazi i podatak o prosečnoj mesečnoj molskoj frakciji $CO_2$ određenoj iz dnevnih proseka. Meseci koji nedostaju su označeni sa -99.99. Inicijalni skup podataka sadrži sledeće kolone: Date, *Decimal Date*, *Average*, *Interpolated*, *Trend*, *Number of Days*, a podatak koji je od interesa za ovaj rad je:

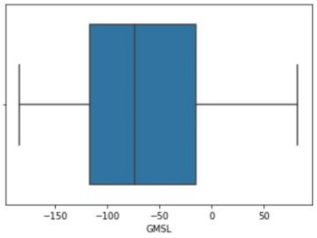
- *Average* - prosečna koncentracija $CO_2$ na mesečnom nivou.

Uklonjene su kolone koje nisu od interesa i uzeti su u obzir datumi od 1969. do 2013. godine. Pošto skup podataka ima 727 redova, a broj jedinstvenih datuma je takođe 727, zaključeno je da ne postoje duplikati.

Skup podataka [13] koji sadrži promenu nivoa mora od 1880. do 2013. godine, kreiran je pomoću informacija dobijenih iz satelitskih snimaka. Bitni podaci iskorišćeni u ovom radu su:

- *Date* - datum merenja i
- GMSL - globalna srednja vrednost nivoa mora.

Inicijalni skup podataka sadrži atribute *Time*, GMSL, GMSL *uncertainty*. Uklonjena je kolona GMSL *uncertainty* i preimenovana kolona *Time* u *Date*. Izvršena je provera da li postoje nedostajuće vrednosti i daljom analizom zaključeno je da ne postoje. Takođe, kao što se može videti sa slike 2, utvrđeno je da ne postoje ni autlajeri. Pošto skup podataka ima 1608 redova, a broj jedinstvenih datuma je takođe 1608, zaključuno je da ne postoje duplikati i uzeti su u obzir datumi od 1969. do 2013. godine.



Slika 2. Grafiki prikaz autlajera za GMSL

## IV. METODOLOGIJA

Metodologija ovog rešenja može se podeliti u tri celine: analiza i integracija podataka za porast nivoa mora, izbor i optimizacija modela za predviđanje da li se desio porast nivoa mora u odnosu na prethodni mesec i predviđanje porasta nivoa mora korišćenjem *Time Series*.

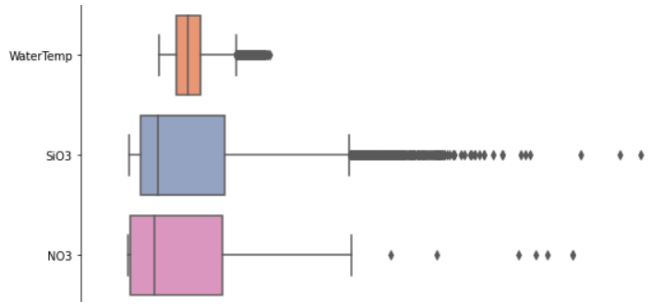
### *A. Analiza i integracija podataka za porast nivoa mora*

Kako su u prethodnom poglavlju detaljno opisani svi skupovi podataka, kao i značaj relevantnih atributa, u ovom delu će biti reči o samoj analizi i integraciji podataka. Svaki od prethodno navedenih skupova podataka detaljno je analiziran i procesiran, nakon čega su relevantni podaci sačuvani u novi skup podataka za kasniju intergraciju. Treba napomenuti da je podatak o datumu merenja od izuzetnog značaja, s obzirom da se integracija skupova podataka vršila na osnovu ovog atributa.

Prilikom analize i obrade podataka, najpre su iz učitanog skupa podataka uklonjeni atributi koji nisu od interesa za ovaj rad, nakon čega su izvršene neke od standarnih operacija za proveru nedostajućih vrednosti, duplikata, kao i za proveru postojanja autlajera. Ovi koraci su primenjeni nad svakim

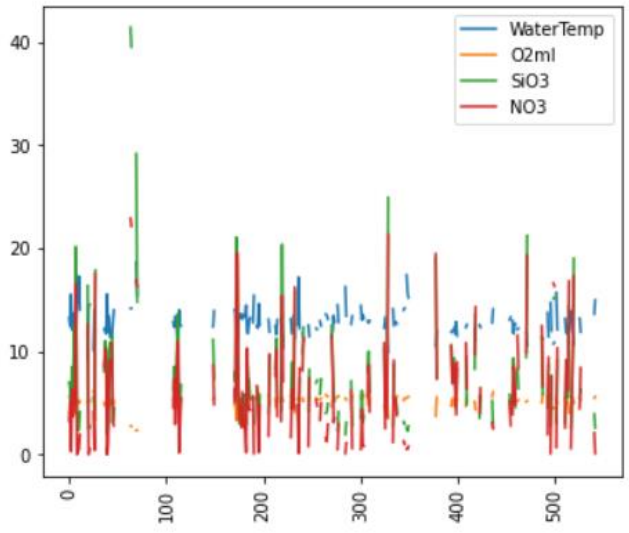
skupom podataka, dok je većina zahtevala dodatu analizu i obradu podataka.

U okviru skupa podataka koji sadrži informacije o vodama, uočeno je postojanje autlajera koji su prikazani na slici 3 Empirijskom metodom, došlo se do zaključka da bi njegove vrednosti trebalo da budu naknadno otklonjene.



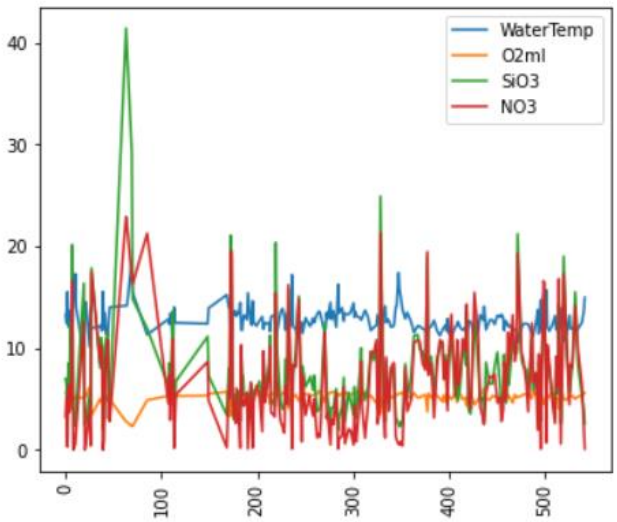

Slika 3. Prikaz uočenih autlajera

U okviru ovog skupa podataka, uočen je veliki broj nedostajućih vrednosti za atribute koji se od značaja u ovom radu. Na slici 4 može se videti njihov grafički prikaz u vremenskom periodu od 1969. do 2013. godine.




Slika 4. Grafički prikaz podataka o vodama sa nedostajućim vrednostima

Za popunjavanje prethodno prikazanih nedostajućih vrednosti iskorišćena je interpolacija, gde je kao metod izabrana polinominalna funkcija koja se uz dodatnu konfiguraciju pokazala kao najbolje rešenje u odnosu na sve ispitane tipove interpolacije, u koje spadaju: *near*, *slinear*, *quadratic* i *barycentric*. Vrednosti podataka nakon njene primene prikazani su na slici 5.




Slika 5. Grafički prikaz podataka o vodama nakon primene polinominalne interpolacije za popunjavanje nedostajućih vrednosti

U okviru skupa podataka koji sadrži informacije o ukupnoj površini morskog leda, pored standarnih operacija izvršeno je grupisanje podataka po datumu merenja za severnu i južnu hemisferu korišćenjem *mean* metode, dok su nedostajuće vrednosti popunjene korišćenjem *Curve Fitting* [9], matematičke funkcije koja se pokazala kao najbolje rešenje.

Skup podataka koji sadrži informacije o promenama nivoa mora je od izuzetnog značaja, s obzirom da je atribut koji predstavlja globalnu srednju vrednost nivoa mora iskorišćen za formiranje labele u konačnom skupu podataka. Vrednosti koje ova labela uzima su:

- 0 - nije se desio porast nivoa mora u odnosu na prethodni mesec i
- 1 - desio se porast nivoa mora u odnosu na prethodni mesec.

Za razmatrane godine 256 puta se desilo smanjenje, a 284 puta povećanje nivoa mora u odnosu na prethodni mesec.

Vrednosti GMSL atributa su takođe iskorišćene prilikom predviđanja porasta nivoa mora korišćenjem *Time Series*.

### B. *Predviđanje da li se desio porasta nivoa mora u odnosu na prethodni mesec*

Za rešavanje ovog problema najpre je isprobano nekoliko jednostavnih modela kao što su *Naive-Bayes* i SVM, nakon čega su iskorišćeni *Random Forest* i *Bagging* modeli sa i bez optimizacije parametara, dok je najviše vremena posvećeno *XGBoost* modelu i njegovoj optimizaciji. Ulaz u svaki od ovih modela predstavlja prethodno opisani konačni skup podataka, dok izlaz iz modela predstavlja vrednost da li se desio porast nivoa mora u odnosu na prethodni mesec.

Prilikom korišćenja *Naive-Bayes* i SVM modela početni skup podataka podeljen je na trening i test u odnosu 75:25 [6], nakon čega je izvršena normalizacija podataka primenom formule (1):

$$X' = \frac{X - Xmin}{Xmax - Xmin} \quad (1)$$

gde je:

- Xmin – minimalna vrednost obeležja na trening skupu i
- Xmax – maksimalna vrednost obeležja na trening skupu.

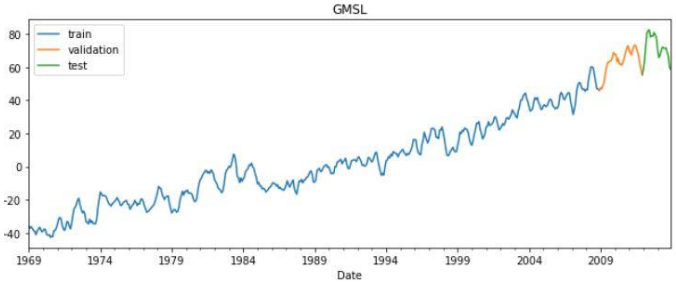
Normalizovani podaci su iskorišćeni za treniranje modela, nakon čega je izvršena predikcija nad testnim podacima. Oba modela su dala veoma slične rezultate.

*Random Forest* i *Bagging* modeli primenjeni su u cilju poboljšanja rezultata. Početni skup podataka najpre je podeljen u odnosu 75:25 [6], nakon čega se daljim isprobavanjem došlo do zaključka da se odnos 70:30 pokazao kao najbolji izbor. Za početak, isproban je *Random Forest* klasifikator bez podešavanja parametara, čiji su rezultati bili nešto bolji u odnosu na *Naive-Bayes* i SVM. Prilikom optimizacije parametara korišćen je *RandomizedSearchCV* kako bi se pronašle najoptimalnije vrednosti parametara za broj estimatora, maksimalnu dubinu stabla i maksimalne karakteristike stabla. Za broj estimatora definisane su vrednosti od 100 do 1000 sa korakom 10, dok su za maksimalnu dubinu definisane vrednosti od 10 do 100 sa korakom 5. Maksimalne karakteristike mogu uzeti vrednosti auto, sqrt ili $\log_2$. Nakon primene najoptimalnijih parametara, rezultati su u maloj meri poboljšani. *Bagging* model dao je slične rezultate sa optimizacijom parametara. Najpre je kao bazni estimator definisan *DecisionTreeClassifier* gde je za kriterijum izabrana entropija koja služi za dobijanje informacija, a za *random state* dodeljena je vrednost 9, dok je za maksimalnu dubinu dodeljena vrednost 20. Pored baznog estimatora, definisan je broj estimatora čija je vrednost 95, broj uzoraka sa vrednošću 0.9 i *random state* sa vrednošću 9.

*XGBoost* je poslednji model koji je korišćen za predviđanje da li se desio porast nivoa mora u odnosu na prethodni mesec. Nakon podele na obučavajući i testni skupove, podaci su transformisani u *DMatrix* objekte kako bi se izvršilo obučavanje *XGBoost* modela. Za proveru tačnosti modela, korišćena je srednja apsolutna vrednost greške. Prilikom prvog obučavanja modela definisani su parametri za binarnu klasifikaciju zajedno sa atributom *objective* kome je dodeljena vrednost *binary:logistic*. Rezultati su bili nešto lošiji u odnosu na prethodno primenjene modele, nakon čega je izvršena unakrsna validacija radi podešavanja hiper-parametara modela. Za pronalaženje najoptimalnijih vrednosti parametara korišćen je *RandomizedSearchCV* gde je kao estimator definisan *XGBRegressor*. Svakom od atributa dodeljen je niz mogućih vrednosti koje su formirane na osnovu nekoliko primera koji se bave binarnom klasifikacijom. Kao rezultat dobijene su sledeće vrednosti parametara: *subsample*: 1.0, *min_child_weight*: 6, *max_depth*: 3, *learning_rate*: 0.01, *eta*: 0.2, *colsample_bytree*: 0.8, nakon čega je izvršeno ponovno treniranje modela sa optimizovanim parametrima. Rezultati su u ovom slučaju poboljšani u odnosu na prvobitno treniranje modela, ali su i dalje bili dosta slični rezultatima koje daju *RandomForest* i *Bagging* modeli nakon optimizacije parametara. Takođe, određena je značajnost atributa po *XGBoost*-u prema kom koncentracija $O_2$ i $NO_3$ imaju najveći uticaj, nakon čega slede temperature zemlje i okeana što je i bilo očekivano.

### C. *Predviđanje porasta nivoa mora korišćenjem Time Series*

U ovom segmentu rada, predviđan je tačan porast globalne srednje vrednosti nivoa mora rešavanjem *Time Series* problema. Vrednosti prethodno kreirane labele, koje označavaju povećanje i smanjenje nivora mora, nisu više od značaja, već je skup podataka proširen vrednostima GMSL atributa. Obzirom na datume merenja faktora koji utiču na globalno zagrevanje, kao korak za predviđanje *Time Series*-a odabran je interval od mesec dana. Skup podataka podeljen je na obučavajući skup u rasponu od 01.1969. do 12.2008. godine, validacioni skup od 12.2008. do 12.2011. godine i testni skup u rasponu od 12.2011. do 12.2013. godine. Ovakav način podele skupa podataka proporcionalno odgovara podeli koja je prikazana u referentnom radu ,,Forecasting time series with gradient boosting: Skforecast, XGBoost, LightGBM y CatBoost“ [15] i prikazan je na slici 6.




Slika 6. Grafički prikaz podele skupa podataka

Vrednosti GMSL-a biće predviđane u ovom delu rada, dok će ostali podaci poslužiti za njegovo predviđanje.

Za predviđanje korišćen je *ForecasterAutoreg*, pri čemu je za regresor odabran *XGBRegressor*, a empirijskim putem izveden je zaključak da predviđanje sa korakom 8 daje najbolje rezultate. Prilikom podešavanja hiper-parametara, isprobavanjem i zadavanjem većeg koraka pri obučavanju modela, dobijeni su sve lošiji rezultati. Model je kao najbolje parametre izdvojio vrednosti 0.1 za *learning_rate*, 10 za *max_depth* i 500 za *n_estimators*. Takođe, izdvojio je niz za *lags* od 1 do 10, sa korakom 1.

Kao što se moglo i pretpostaviti, vremenski koraci najviše utiču na previđanje modela, a zatim sledi debljina leda, temperatura i emisija ugljen-dioksida.

## V. REZULTATI I DISKUSIJA

Rezultati ovog rada biće podeljeni u dva odvojena segmenta, a koja predstavljaju dva odvojena pristupa pri rešavanju ovog problema. Prvi segment odnosiće se na rezultate dobijene kroz različite obučene modele, sa naglaskom na najbolje dobijenom modelu. Rezultati *Time Series* problema, činiće drugi segment ovog poglavlja.

### A. *Rezultati obučenih modela*

Nakon obučavanja modela, bilo je potreno odrediti njihovu preciznost, grešku ili određenu meru evaluacije. Kao metod evaluacije korišćeni su preciznost (*precision*), odziv (*recall*) i F mera (*F1-measure*). F mera je data formulom (2):

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

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

$$F1 - measure = 2 * \frac{precision * recall}{precision + recall} \qquad (2)$$

gde je:

- TP – podatak za koji model kaže da se porast nivoa mora desio i da taj podatak zapravo jeste podatak o porastu nivoa mora,
- FP – podatak za koji model kaže da se porast nivoa mora desio, a da taj podatak zapravo nije podatak o porastu nivoa mora i
- FN – podatak za koji model kaže da se porast nivoa mora nije desio, a da taj podatak zapravo jeste podatak o porastu nivoa mora.

Prvi trenirani modeli bili su *Naive-Bayes* i SVM klasifikator. Ovi modeli trenirani su nad skupom podataka koji je podeljen na obučavajući i testni skup u razmeri 75:25 [6]. Učinak tačnosti koji modeli pokazuju su identični i iznose svega 54%. U TABELA I prikazani su rezultati prilikom evaluacije modela, pri čemu se nulom u redu označava da se porast mora nije desio, a jedinicom da jeste, dok kolone sačinjavaju preciznost, odziv i F mera. Ovakva struktura tabele ponavljaće se u ovom poglavlju. Iz ovog rezultata može se uočiti da, zbog globalnog rasta nivoa mora, modeli bolje predviđaju povaćanje u odnosu na prethodni mesec nego opadanje.

TABELA I REZULTAT EVALUACIJE – *NAIVE BAYES* I SVM KLASIFIKATOR

| ***Rezultati*** | ***Preciznost*** | ***Odziv*** | ***F mera*** |
|---|---|---|---|
| 0 | 0.47 | 0.41 | 0.44 |
| 1 | 0.60 | 0.65 | 0.62 |

Naredni trenirani model je *XGBoost*, koji je treniran nad istim skupom podataka kao i prethodni modeli i sa identičnom razmerom obučavajućeg i testnog skupa. Osnovna srednja apsolutna greška iznosila je 50%. Prilikom prve iteracije modela, a bez optimizacije parametara, greška je smanjena na 48%. U TABELA II prikazana je F mera. Dobijeni rezultati su prilično lošiji za razliku od rezultata dobijenih od prethodna dva modela.

TABELA II REZULTAT EVALUACIJE – *XGBOOST*

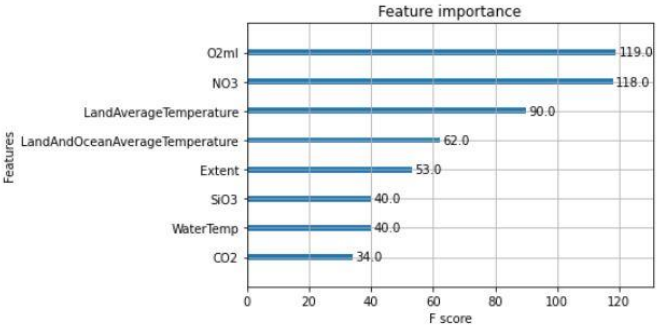
| ***Rezultati*** | ***Preciznost*** | ***Odziv*** | ***F mera*** |
|---|---|---|---|
| 0 | 0.42 | 0.52 | 0.47 |
| 1 | 0.56 | 0.47 | 0.51 |

Nakon podešavanja hiper-parametara modela, rezultati su se poboljšali i to za 10% za predviđanje da li se desio porast nivoa mora. U TABELA III prikazana je F mera.

TABELA III REZULTAT EVALUACIJE – *XGBOOST NAKON PODEŠAVANJA HIPER-PARAMETARA*

| ***Rezultati*** | ***Preciznost*** | ***Odziv*** | ***F mera*** |
|---|---|---|---|
| 0 | 0.51 | 0.53 | 0.52 |
| 1 | 0.64 | 0.61 | 0.62 |

Na slici 7, prikazana je značajnost atributa po *XGBoost*-u. Može se uočiti da najveći uticaj na porast nivoa mora imaju kiseonik, koji je sastavni deo vode, nitrat ($NO_3$), kao i temperatura kopna. Nitrat je element koji je sastavni deo đubriva i koristi se u poljoprivredi. U kombinaciji sa ugljen-dioksidom dokazano doprinosi povećanju globalnog zagrevanja [7].




Slika 7. Značajnost atributa za *XGBoost*

Algoritam *Random Forest* nije doprineo boljim rezultatima. U TABELA IV dati su rezultati F mere koja se drastično ne razlikuju od ostalih, prethodno prikazanih. Međutim, nakon podešavanja hiper-parametara modela, dobili smo najbolje predviđanje za situaciju kada gledamo smanjenje nivoa mora i ovi rezultati su prikazani u TABELA V.

TABELA IV REZULTAT EVALUACIJE – *RANDOM FOREST*

| ***Rezultati*** | ***Preciznost*** | ***Odziv*** | ***F mera*** |
|---|---|---|---|
| 0 | 0.60 | 0.48 | 0.53 |
| 1 | 0.60 | 0.71 | 0.65 |

TABELA V REZULTAT EVALUACIJE – *RANDOM FOREST NAKON PODEŠAVANJA HIPER-PARAMETARA*

| ***Rezultati*** | ***Preciznost*** | ***Odziv*** | ***F mera*** |
|---|---|---|---|
| 0 | 0.60 | 0.53 | 0.57 |
| 1 | 0.62 | 0.68 | 0.65 |

Poslednji model za koji su razmatrani rezultati predikcije bio je *Bagging*. U TABELA VI dati su rezultati F mere nakon

podešavanja hiper-parametara, dok su rezultati sa osnovnim podešavanjima dali identične rezultate kao i *Random Forest*.

TABELA VI REZULTAT EVALUACIJE – *BAGGING NAKON PODEŠAVANJA HIPER-PARAMETARA*

| ***Rezultati*** | ***Preciznost*** | ***Odziv*** | ***F mera*** |
|---|---|---|---|
| 0 | 0.62 | 0.49 | 0.55 |
| 1 | 0.61 | 0.72 | 0.67 |

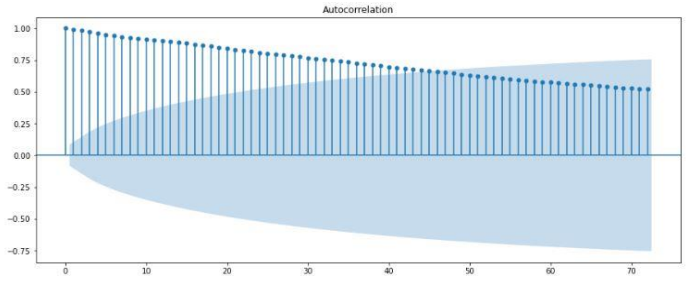
Nakon proučenih, dosta loših, rezultata svih obučenih modela, došlo se do zaključka da svi daju približno iste rezultate bez obzira na podešavanje hiper-parametara. Pomoću *XGBoost* modela, dobijena je realnija slika, koja pokazuje da najznačajni atributi zaista jesu ispravni i podudaraju se sa ključnim faktorima globalnog zagrevanja. Jedno od mogućih rešenja jeste da merenje faktora koji utiču na globalno zagrevanje bude kontinualno, kako bi skup podataka postao složeniji i ujednačeniji i time model tačnije predviđao.

U narednom segmentu ovog poglavlja, prikazani su rezultati pristupa pomoću *Time Series*-a.

### *B. Time Series rezultati*

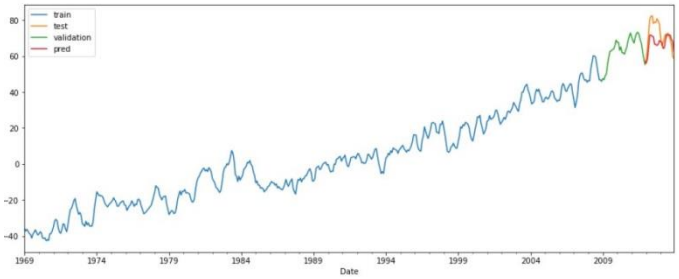
Dosadašnji cilj, u kom su faktori koji utiču na globalno zagrevanje pokušali da odgonetnu da li je postajalo povećanje nivoa mora, nije naišao na uspeh. Međutim, *Time Series* pristup doprineo je mnogo boljim rezultatima i tačnijim analizama. U ovom segmentu, fokus je na predikciji tačne vrednosti srednjeg globalnog nivoa mora, uz pomoć različitih činilaca i vremenskog koraka.

Na osnovu grafika korelacije na slici 8, može se doći do zaključka da su vrednosti GMSL-a povezane, bez naglih oscilacija.




Slika 8. Autokorelacija GMSL

Nakon predviđanja tačne vrednosti nivoa mora, pored vremenskih koraka, faktori koji najviše doprinose krajnjim rezultatima jesu debiljna leda, temperatura i molekule kiseonika u vodi. Ove vrednosti su slične onima koje je izdvojio i *XGBoost* model. Na slici 9, dat je prikaz originalnih i previđenih vrednosti. Uočeno je da model lakše predviđa opadanje u odnosu na porast nivoa mora.




Slika 9. Predikcija *Time Series*-a prikazana crvenom linijom

U TABELA VII i TABELA VIII izdvojena su po tri meseca, za koje se predviđene i originalne vrednosti najviše i najmanje razlikuju.

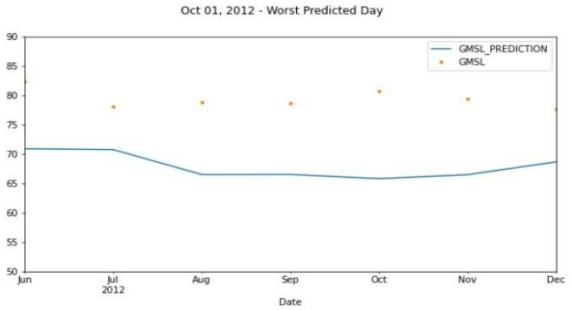
TABELA VII. REZULTAT *TIME SERIES*-A – *NAJGORI MESECI*

| ***Dan*** | ***GMSL*** | ***Previđeni GMSl*** | ***Apsolutna greška*** |
|---|---|---|---|
| 10-2012 | 80.8 | 65.8 | 15.0 |
| 11-2012 | 79.4 | 66.5 | 12.9 |
| 08-2012 | 78.8 | 66.5 | 12.3 |

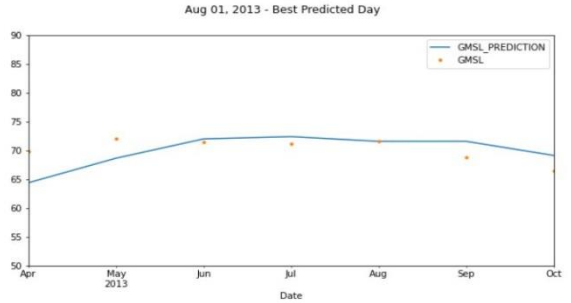
TABELA VIII. REZULTAT *TIME SERIES*-A – *NAJBOLJI MESECI*

| ***Dan*** | ***GMSL*** | ***Previđeni GMSl*** | ***Apsolutna greška*** |
|---|---|---|---|
| 08-2013 | 71.6 | 71.57 | 0.03 |
| 06-2013 | 71.5 | 72 | 0.5 |
| 07-2013 | 71.2 | 72.4 | 1.2 |

Na slikama 10 i 11 data je grafička predstava najboljeg i najgoreg previđanog meseca.




Slika 10. Najgore predviđani mesec




Slika 11. Najbolje predviđani mesec

Analizom dobijenih rezultata, može se uočiti da se dogodio nagli skok srednje vrednosti nivoa mora u najlošijem previđenom mesecu. Paralelno sa time, model skoro potpuno precizno predviđa opadanje nivoa mora. Ukoliko bi se posedovali podaci na dnevnom, a ne mesečnom nivou, ova predviđanja bi postala još tačnija, obzirom da se *Time Series* pokazao kao najbolji pristup za rešavanje ovog problema.

Nivo mora u 2012. godini poseduje nagli skok, a koji se može videti na slici 9. Ujedno, *Time Series* najviše greši za ovaj period, ali i prikazuje da je temperatura jedan od faktora koji najviše utiču na porast srednje vrednosti nivoa mora. Ovi rezultati se mogu opravdati činjenicom da je 2012. godina, bila najtoplija godina od kako se mere temperature na Zemlji [8] i to sa rekordno visokom temperaturom tokom jeseni. Ovaj podatak se poklapa sa činjenicom da je dobijen najgori rezultat za oktobar.

## VI. Zaključak

U ovom radu predstavljeno je rešenje predviđanja porasta nivoa mora koje je posledica sve većeg globalnog zagrevanja. Kako bi se još više probudila svest kod ljudi o ovom problemu i dublje analizirali faktori koji na njega utiču, predstavljena su dva odvojena pristupa koja daju osnov i motivaciju za njihovo dalje unapređenje.

Najpre su prikupljeni relevantni podaci integracijom nekoliko skupova podataka i čiji zajednički atribut predstavlja datum merenja koje je vršeno na mesečnom nivou. Prvo rešenje predstavlja odgovor na pitanje da li se desilo povećanje nivoa mora u odnosu na prethodni mesec i koji faktori na njega najviše utiču. Modeli koji su iskorišćeni za rešavanje ovog problema su: SVM, *Naive-Bayes*, *Random Forest*, *XGBoost* i *Bagging*. Poređene su performanse svakog modela nad konačnim skupom podataka, dok je za evaluaciju tačnosti modela izabrana F mera. Svi obučeni modeli su i nakon optimizacije hiper-parametara dali približno slične rezultate, dok se *XGBoost* pokazao kao najkorisniji. Drugo rešenje, predstavlja predviđanje tačne vrednosti srednjeg globalnog nivoa mora, korišćenjem *Time Series*-a, a koji pored faktora koji utiču na porast nivoa mora, koristi i vremenski korak. Ovaj pristup se ispostavio kao koristan za dalje unapređenje rešenja ovog problema.

Glavni nedostatak ovog rada ogleda se u maloj količini skupova podataka koji imaju isti vremenski raspon tokom kog su prikupljani relevantni atributi koji utiču na porast nivoa mora. Različiti skupovi ne poseduju konstantno merene vrednosti i time prouzrukuju ogroman broj nedostajućih podataka. Ukoliko bi se nastavio trend konzistentnog prikupljanja podataka, modeli bi preciznije predviđali i samim tim dali tačnije rezultate. Ovim bi se izdvojili najveći činioci koji utiču na porast nivoa mora i stekao bi se uslov za njihovu detaljniju analizu.

Planovi za dalji rad:

- *Explainability* najbolje dobijenog modela,
- ponovno obučavanje modela sa podacima prikupljenih dodatnim merenjima radi preciznijih rezultata,
- detaljnija analiza faktora za koje model predviđa da najviše utiču na porast nivoa mora,
- dodavanje dodatnih faktora koji bi potencijalno uticali na preciznost modela nad velikim skupom podataka i
- korišćenje *Time Series*-a sa dnevnim korakom predviđanja, ukoliko bi se svakodnevno beležile vrednosti iz skupova podataka.

## Literatura


[1] Al. Balogun, N. Adebisi, “Geomatics, Natural Hazards and Risk”, vol. 12, 2021.

[2] Harvey Zheng, “Analysis of Global Warming Using Machine Learning”, vol. 7, no. 3 2018.

[3] Veronica Nieves, Christina Radin, Gustau Camps-Valls, “Predicting regional coastal sea level changes with machine learning”, 2021.

[4] Magnus Hieronymos, Jenny Hieronymos, Frederik Hieronymos, “On the Application of Machine Learning Techniques to Regression Problems in Seal Level Studies”, Sep. 2019.

[5] Christophe Bonneuil, Pierre-Louis Choquet, Benjamin Franta, “Early warning and emerging accountability: Total’s responses to global warming, 1971-2021”, vol. 71, Nov. 2021

[6] Geoffrey I. Webb, Janice R. Boughton, Zhihai Wang, “Not So Naive Bayes: Aggregating One-Dependence Estimators”, January 2005

[7] Xiaoyu Li, Shulan Cheng, Huajun Fang, Guirui Yu, Xushen Dang, Minjie Xu, Lei Wang, Gaoyue Si, Jing Geng, Shun He, “The contrasting effects of deposited NH4 and NO3 on soil CO2, CH4 and NO2 fluxes in subtropical plantation, southern China”, December 2015

[8] [Online] National Temperature and Precipitation Analysis. Available: https://www.ncei.noaa.gov/access/monitoring/monthly-report/national/201213 [Accessed: 21-Apr-2022]

[9] [Online] Curve Fitting. Available: https://machinelearningmastery.com/curve-fitting-with-python/ [Accessed: 21-Apr-2022]

[10] [Online] Water data. Available: https://www.kaggle.com/datasets/mathsian/water-temperature [Accessed: 21-Apr-2022]

[11] [Online] Climate Change: Earth Surface Temperature Data. Available: https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv [Accessed: 21-Apr-2022]

[12] [Online] CO2 PPM – Trends in Atmospheric Carbon Dioxide. Available: https://datahub.io/core/co2-ppm [Accessed: 21-Apr-2022]

[13] [Online] Global Average Absolute Sea Level Change, 1880-2014. Available: https://datahub.io/core/sea-level-rise [Accessed: 21-Apr-2022]

[14] [Online] Daily Sea Ice Extent Data. Available: https://www.kaggle.com/datasets/nsidcorg/daily-sea-ice-extent-data [Accessed: 21-Apr-2022]

[15] [Online] Forecasting time series with gradient boosting: Skforecast, XGBoost, LightGBM y CatBoost by Joaquín Amat Rodrigo. Available: https://www.cienciadedatos.net/documentos/py39-forecasting-time-series-with-skforecast-xgboost-lightgbm-catboost.html [Accessed: 21-Apr-2022]