Predlog projekta iz predmeta

„Sistemi za istraživanje i analizu podatka“

Ovaj dokument sadrži kratak opis definicije projekta, kao i motivaciju za odabranu temu. Zatim je navedena relevantna literatura u kojoj su ukratko prezentovani naučni radovi sa opisom srodnih problema i načinom njihovog rešavanja. Predstavljen je inicijalni skup podataka, metodologija rada i metod evaluacije. Na samom kraju, opisani su softver i plan rada.

1. Definicija projekta

Glavni cilj ovog projekta ogleda se u analizi različitih faktora koji utiču na globalno zagrevanje i samim tim na konstantan porast nivoa mora. Globalno zagrevanje predstavlja povećanje prosečne temperature zemljine atmosfere i okeana, čime je uslovljen i porast nivoa mora na koji utiču razni faktori kao što su: temperatura na površini, otapanja glečera, salinitet, površinska gustina mora i drugo. Ideja je najpre analizirati promenu temperature kopna i okeana tokom godina, njen uticaj na srednju kumulativnu masu glečera. Na kraju, biće izvršena tendencija rasta ili smanjenja nivoa mora u narednih nekoliko godina.

2. Motivacija

Porast nivoa mora je globalno ekološko pitanje i pretnja po obalne zajednice koje se obično karakterišu velikom gustinom stanovništva i infrastrukture. Mnogobrojni faktori utiču na ove promene, što je u velikoj meri uslovljeno globalnim zagrevanjem čije posledice mogu biti katastrofalne. Promena obale, sve veća učestalost i intenzitet poplava i podzemnih voda su samo neki od fizičkih uticaja, dok društveno-ekonomske posledice uključuju degradaciju kvaliteta vode, oštećenje infrastrukture, iscrpljenje poljoprivrednih resursa i drugo. Stoga je procena porasta nivoa mora, kao i detaljna analiza faktora kojima su ove promene uslovljene od velike važnosti za održiv razvoj priobalnih zajednica, kao i za ostatak čovečanstva.

3. Relevantna literatura

[1] Al. Balogun, N. Adebisi - Geomatics, Natural Hazards and Risk (2021) *Sea level prediction using ARIMA, SVR and LSTM neural network: accessing the impact of enseble Ocean-Atmospheric processes on models' accuracy.*
https://www.tandfonline.com/doi/pdf/10.1080/19475705.2021.1887372
Tema rada: Ova studija ima za cilj da integriše širok spektar okeansko-atmosferskih varijabli za predviđanje varijacija nivoa mora duž obale Zapadnog poluostrva Malezije koristeći mašinsko učenje i tehnike dubokog učenja.
Metodologija: Izvršeno je više scenarija različitih kombinacija varijabli korišćenjem ARIMA, SVR i LSTM neuronske mreže.
Podaci: Podaci o moru uključujući anomaliju nivoa mora (SLA), temperaturu površine mora (SST), salinitet površine mora (SSS) i gustina površine mora (SSD) dobijeni su od Copernicus

službe za praćenje morske sredine (CMEMS), kao i atmosferski podaci uključujući padavine, ukupnu oblačnost i brzinu vetra. Svi podaci ucrtani su u mrežu srednjih mesečnih skupova podataka između januara 1993. i oktobra 2019. Korišćeno je 7 mernih stanica koje se nalaze na strateškim lokacijama oko obala Zapadnog poluostrva Malezije.

Evaluacija: Usvojen je standardni odnos 70:30 za skupove podataka za obuku i testiranje. U ovoj studiji, koeficijent korelacije R se koristi za procenu tačnosti predviđanja modela.

Rezultati: Poređenje modela obučenih na okeanskim varijablama sa modelima obučenim na atmosferskim varijablama pokazuje da atmosferski procesi imaju veći uticaj na predviđanje modela nego procesi okeana. Obuka sa kombinovanim okeansko-atmosferskim varijablama poboljšala je predviđanja modela na svih 7 stanica. Srednja vrednost R tačnosti LSTM, SVR i ARIMA modela sa optimalnim performansama na svim stanicama je 0.85, 0.74 i 0.71.

Zaključak: Analizom različitih scenarija kombinacijom okeansko-atmosferskih varijabli ustanovljeno je da atmosferski procesi imaju veći uticaj na predviđanje modela nego procesi okeana, što ćemo uzeti u obzir prilikom obrade skupova podataka i izbora relevantnih atributa. Za procenu tačnosti predviđanja modela koeficijent korelacije R pokazao kao dobar izbor.

[2] Harvey Zheng (2018) *Analysis of Global Warming Using Machine Learning*
https://www.scirp.org/html/5-2570177_86337.htm

Tema rada: Konstrukcija pouzdanih statističkih modela zasnovanih na ogromnim klimatskim podacima prikupljanih mnogo godina u nazad, kao i precizna identifikacija odnosa između temperature i potencijalnih faktora kao što su koncentracija ugljen-dioksida (CO_2), azot-oksida (N_2O) i metana (CH_4). Izgradnja najsavremenijeg modela za verifikaciju zagrevanja zemlje i identifikovanja faktora koji doprinose globalnom zagrevanju.

Metodologija: Algoritmi mašinskog učenja primenjeni u ovoj studiji o globalnom zagrevanju su *Random Forest*, regresija vektora podrške (SVR), *Lasso* i linearna regresija.

Podaci: Iz raznih javnih baza kao što su „Nacionalna uprava za okeane i atmosferu“ i „Agencija za zaštitu životne sredine Sjedinjenih Država“, prikupljeni su podaci o CO_2, N_2O i CH_4. Precizni podaci o prosečnoj temperaturi u poslednjih 100 godina dobijeni su od „Lawrence Berkeley National Lab“. S obzirom da nisu svi podaci međusobno usklađeni, korišćena je linearna interpolacija za njihovo usklađivanje i efikasniju obradu.

Evaluacija: Prikupljeni podaci nasumično su podeljeni u dva jednaka skupa, jedan za trening i jedan za testiranje. Tokom procesa obuke, korišćena je osmostruka unakrsna validacija za traženje odgovarajućih hiper-parametara i sprečavanje *overfitting-a* modela tokom treninga. Zatim su upoređena tri različita algoritma mašinskog učenja (*Random Forest*, *Lasso* i SVR) gde su parametri podešeni da odgovaraju podacima i generišu tačne rezultate obuke.

Rezultati: *Random Forest* nadmašio je ostale korišćene algoritme mašinskog učenja. Vizuelno je precizniji i stvara najtačniji model za predviđanje temperaturnih razlika. Pokazao je da na promenu temperature najviše utiče CO_2, zatim CH_4 i na kraju N_2O.

Zaključak: Iz ove studije zaključili smo da se *Random Forest* pokazao kao najtačniji model i da CO_2 najviše utiče na promenu temperature što ćemo dodatno analizirati. Takođe, uvažićemo predlog za primenu drugih algoritama poput *XGBoost* radi dobijanja preciznijih rezultata.

[3] Veronica Nieves, Christina Radin, Gustau Camps-Valls (2021) *Predicting regional coastal sea level changes with machine learning*
https://www.nature.com/articles/s41598-021-87460-z

Tema rada: Napredne statističke analize, uključujući metode mašinskog učenja (ML), mogu pružiti koristan uvid u promenu nivoa mora. Glavni cilj ovog rada je potreba za određivanjem tendecija povećanja i smanjenja nivoa mora u narednim godinama uz pomoć mašinskog učenja. Predloženi pristup kombinuje promene temperature površine okeana, u regionima otvorenog mora i ML tehnike za procenu varijabilnosti obalnog nivoa u ovim regionima.
Metodologija: U radu je korišćen model regresije Gausovog procesa (GP) i rekurentna neuronska mreža (RNN) sa jedinicama dugotrajne memorije (LTSM).
Podaci: Korišćeni podaci se baziraju na odnosu između procene prosečne vertikalne temperaturne anomalije (MTA) i dubinske integrisane temperature tj. sadržaja toplote (OHC) u regionima otvorenog mora i priobalnih anomalija nivoa mora (SLA). Analizirao se uticaj otvorenog okeana, sa njegovom promenom prosečne temperature do 700m dubine, na obalu.
Evaluacija: Skup podataka koristeći GP je podeljen na trening skup 88.5% (1993-2015) i test skup 11.5% (2016-2018), što se pokazalo kao dobro rešenje ne samo za identifikaciju promene varijabilnosti regionalnog obalnog nivoa mora, već bi se moglo iskoristiti i za popunjavanje praznina u podacima. Za obučavanje modela kod RNN uzet je deo poslednjih godina za trening skup, tri godine za testiranje predviđanja i jedna godina za validaciju.
Rezultati: Uspešno su iskorišćene temperature gornjih slojeva otvorenog okeana pri proceni nivoa mora na obalnim mestima. Data su približno tačna predviđanja tendencije nivoa mora u razmaku od jedne do tri godine.
Zaključak: Skup podataka koji poseduje promenu temperature okeana kroz godine, ispostavio se kao dobar parametar pri proceni tendencije nivoa mora. Podela skupa podataka u odnosu na mernu godinu, biće uzeta u razmatranje i u našem radu. U radu 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 ćemo uvažiti.

4. Skup podataka

Nakon obrade skupa podataka, kao i odabira bitnih atributa, grupisaćemo podatke u jedan skup na osnovu zajedničkog atributa koji predstavlja datum registracije podataka.

- https://datahub.io/core/sea-level-rise Ovaj skup podataka sadrži promenu nivoa mora od 1880. do 2013. godine primećenu satelitima. Podaci koji su od interesa za naš rad su globalna srednja vrednost nivoa mora (GMSL) i datum merenja.
- https://www.kaggle.com/nsidcorg/daily-sea-ice-extent-data Ukupna površina morskog leda za vremenski period od 1978. do 2015 godine. Za naše potrebe iskoristićemo atribut *Extent* koji predstavlja ukupnu površinu morskog leda izraženu u jedinici 10^6 kvadratnih kilometara.
- https://www.kaggle.com/mathsian/water-temperature Podaci u ovom skupu podataka predstavljaju uzorke vode sakupljene iz okeana između 1959. i 2020. godine. Uzećemo u obzir podatke prikupljane na dubinama do 200m. Pored podatka o datumu uzrokovanja, iskoristićemo podatak o temperaturi vode izražen u celzijusima kao i podatak o zasićenosti vode kiseonikom.
- https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv U okviru ovog skupa podataka nalaze se podaci o globalnoj temperaturi zemlje, okeana i kopna. Podaci su sakupljani u periodu od 1750. do 2015. godine. Za naš rad ćemo uzeti u obzir tri relevantna parametra, a to su datum merenja, globalnu prosečnu temperaturu zemljišta i globalnu prosečnu temperaturu kopna i okeana izražene u celzijusima.

- https://datahub.io/core/co2-ppm Za naše potrebe iskoristićemo skup podataka u kome se pored datuma u kom je izvršeno merenje nalazi podatak o prosečnoj mesečnoj molskoj frakciji CO_2 određenoj iz dnevnih proseka. Meseci koji nedostaju su označeni sa -99.99.
- https://datahub.io/core/co2-fossil-global U okviru ovog skupa podataka nalaze se vrednosti globalne emisije $C0_2$ iz fosilnih goriva od 1751. do 2014. godine. Atributi predstavljaju podatke o emisiji $C0_2$ iz gasnog, tečnog, čvrstog goriva, proizvodnje cementa i spaljivanja gasa, dok ćemo za potrebe našeg projekta uzeti u obzir vrednost ukupne emisije $C0_2$ iz fosilnih goriva.

Zbog rezultata dobijenih iz relevantne literature, obeležja na koja ćemo se fokusirati su: GMSL, *Extent*, temperatura zemlje, okeana i kopna, zasićenost vode kiseonikom i emisija izduvnih gasova ($C0_2$). Usled zavisnosti podataka u odnosu na prethodni mesec, koristićemo razliku vrednosti obeležja u odnosu na prethodni mesec.

Ukoliko bude bilo potrebe iskoristićemo dodatne podatke koje se nalaze u narednom skupu *dataset*-ova - https://datahub.io/collections/climate-change

5. Metodologija

Ulaz u klasifikator predstavljaju prethodno navedena obeležja dobijena iz različitih skupova podataka čijom obradom i spajanjem se dobija novi skup podataka prikazan u Tabeli 1. Na osnovu razlike vrednosti GMSL (globalni srednji nivo mora) obeležja u odnosu na prethodni mesec, formiraće se vrednosti labele. Vrednosti u labeli će biti 1 ukoliko se desio porast nivoa mora u odnosu na prethodni mesec, u suprotnom 0. Nad ulaznim podacima prilikom treniranja vršiće se normalizacija podataka primenom formule:

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

gde je:

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

Izlaz iz modela će predstavljati informacija o tome da li se nivo mora povećao u odnosu na prethodni mesec. Treniranje modela vršiće se unakrsnom validacijom, radi podešavanja hiper-parametara modela. Modeli koji će biti korišćeni su: *XGBoost*, *Random Forest*, SVM, *Bagging*.

DateTime	*Extent*	*WaterTemp*	*O2ml*	*LandAvgTemp*	*LandAndOceanAvgTemp*	*CO2mmf*
1969-01	-48.8	15.17	5.74	1.97	13.52	324.00
1969-02	-49.1	11.59	6.71	2.45	13.75	324.42
…	…	…	…	…	…	…
2015-12	58.5	19.13	5.40	5.52	14.77	401.85

Tabela 1. Primer dataset-a koji predstavlja ulaz u klasifikator.

6. Metod evaluacije

Kao metod evaluacije smatramo da je najbolje koristiti preciznost (*precision*), odziv (*recall*) i F meru (*F1-measure*).

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

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

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

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

Evaluacija će se vršiti nad test podacima koji predstavljaju 20% ukupnog skupa podataka, dok će preostalih 80% predstavljati trening skup. Nakon evaluacije biće odrađena detaljna analiza dobijenih rezultata i Explainability najbolje pokazanog modela.

7. Softver

Za implementaciju ovog projekta koristiće se Python programski jezik sa odgovarajućim bibliotekama za ML (scikit-learn, xgb), analizu i obradu podataka (pandas).

8. Plan

- Prikupljanje podataka
- Grupisanje relevantnih podataka (definisano u tački 4)
- Obrada, transformacija i normalizacija podataka (definisano u tački 5)
- Kreiranje modela
- Evaluacija modela
- Analiza dobijenih rezultata

9. Tim

- Stefan Beljić E2 110/2021
- Stefan Savić E2 111/2021
- Stefan Arađanin E2 112/2021