Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/25BE44Savic

ПРЕДИКЦИЈА ПОРАСТА НИВОА МОРА КОРИШЋЕЊЕМ АЛГОРИТАМА МАШИНСКОГ УЧЕЊА

PREDICTION OF SEA LEVEL RISE USING MACHINE LEARNING ALGORITHMS

Стефан Савић, Јелена Сливка, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *Због утицаја које има на планету и на људске животе, глобално загревање је тема која је постала веома значајна у последњих неколико деценија. Пораст нивоа мора представља једну од најозбиљнијих последица глобалног загревања. Главни циљ овог рада је његова предикција. Најпре је вршено прикупљање релевантних података из различитих извора. Након тога, вршено је претпроцесирање и формирање коначног скупа података који представља улаз у систем. Над претпроцесираним подацима тренинг скупа извршено је обучавање више модела међу које спада метод потпорних вектора, Наивни Бајесов модел, метод случајне шуме, Bagging и XGBoost, као и модели временских серија. Сваки од модела је за излаз имао вероватноћу да ли се десио пораст нивоа мора у односу на претходни месец, док су временске серије предвиђале тачне вредности средњег глобалног нивоа мора. Изведен је закључак да главни утицај на пораст нивоа мора представља емисија штетних гасова тј. угљен-диоксида и температуре, како на површини воде, тако и на копну.*

Кључне речи: *GMSL; глобално загревање; ниво мора; XGBoost; временске серије*

Abstract – *Due to the impact it has on the planet and human lives, global warming is a topic that has become highly significant in the last few decades. Rising sea levels represent one of the most serious consequences of global warming. The main objective of this study is its prediction. First, relevant data was collected from various sources and preprocessed. Multiple models were trained on the preprocessed training set - Support Vector Machine, Naive Bayes model, Random Forest, Bagging, XGBoost, and time series models. Each model outputs the probability of a sea level rise compared to the previous month, while the time series predicted the exact values of the average global sea level occurring. The conclusion is that the primary influence on the sea level rise is the emission of harmful gases (i.e., carbon dioxide) and temperature on the water surface and land.*

Keywords: *GMSL; global warming; sea level; XGBoost; time series*

1. УВОД

Глобално загревање, тј., постепени пораст просечне температуре површине Земље, тема је која је постала веома значајна у последњих неколико деценија због утицаја које има на планету и на наше животе. Оно представља озбиљан изазов за човечанство, са потенцијалом да проузрокује низ штетних последица. Једна од најозбиљнијих последица је пораст нивоа мора услед топљења леда са полова и глечера. Ово може имати озбиљне ефекте на приобална подручја и острва, која су изложена ризику од ерозије и поплава.

Проучавање пораста нивоа мора и анализа фактора који томе доприносе од суштинског су значаја за одрживи развој приобалних заједница и за будућност човечанства. Додатни мотив за бављење овим проблемом је и жеља да се спрече штетне последице по људску економију, друштво и здравље, као и потреба да се очува природа и биодиверзитет.

У овом раду представљена су два решења за одређивање пораста нивоа мора:

1. Први проблем је постављен као проблем бинарне класификације – циљ је био предвидети да ли се десило повећање нивоа мора у односу на претходни месец или не. Такође, одређивани су фактори који су највише допринели повећању нивоа мора. Овај проблем решаван је обуком различитих модела машинског учења (метод потпорних вектора, Наивни Бајесов модел, метод случајне шуме, *XGBoost* и *Bagging*), при чему су као улазни подаци коришћени фактори који утичу на глобално загревање.
2. Други решавани проблем је предикција тачне промене средњег глобалног нивоа мора (GMSL). Овај проблем је решаван применом временских серија са кораком од месец дана.

Као најбољи модел за први проблем показао се *XGBoost*, где је вредност *F* мере за предикцију да се није десило повећање нивоа мора у односу на претходни месец, износила 0.52, а да јесте 0.62. Показало се да највећи утицај на пораст нивоа мора имају кисеоник, који је саставни део воде, нитрат, као и температура копна.

Примена модела временских серија се показала као успешан приступ у добијању жељеног предвиђања нивоа мора. Након предвиђања тачне вредности нивоа мора, поред временских корака, фактори који највише доприносе крајњим резултатима јесу дебљина леда, температура и молекули кисеоника у води, што је

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је била др Јелена Сливка, ванр. проф.

слично показао и *XGBoost* модел. Уколико би се поседовали подаци на дневном, а не месечном нивоу, претпоставка је да би ова предвиђања постала тачнија.

У наредном поглављу, биће приказана сродна истраживања. Затим, у трећем поглављу, биће представљена методологија коришћена при решавању овог проблема, као и спроведени експерименти. Након тога, четврто поглавље ће чинити резултати и дискусија, док ће у петом поглављу бити изведен закључак о самом раду и његовом проблему.

2. ПРЕТХОДНА РЕШЕЊА

У свом раду [1], Балогун и Адебиси интегрисали су широк спектар океанско-атмосферских варијабли ради предвиђања варијација нивоа мора помоћу неуронских мрежа. Компарација модела обучених на океанским и атмосферским варијаблама показује да атмосферски процеси имају већи утицај на предвиђање модела. Ипак, боља предвиђања модела добијају се комбинацијом океанско-атмосферских варијабли. Атмосферски подаци који се користе укључују падавине, облачност и брзину ветра, док подаци о мору укључују температуру, салинитет и густину површине мора. С обзиром на релевантност варијабли и експерименталну поставку, рад [1] се показао као значајан за ово истраживање.

Зхенг [2] је проучавао однос између температуре и других потенцијалних фактора попут CO2, N2O, CH4. Користио је конструкцију статистичких модела базираних на великој количини климатских података и многобројне алгоритме машинског учења као што су: метод случајне шуме, линеарна и регресија вектора подршке и LASSO. Линеарна интерполација коришћена је за усклађивање и ефикаснију обраду података. Током процеса обуке, коришћена је унакрсна валидација за тражење одговарајућих хипер-параметара модела. Потом, упоређена су три различита алгоритма машинског учења (метод случајне шуме, LASSO и SVR). Метод случајне шуме се издвојио као најпрецизнији и он сугерише да највећи утицај на промену температуре има CO2. По узору на рад [2], одлучено је да један од атрибута буде управо CO2, а метод случајне шуме један од коришћених алгоритама. У складу са препорукама аутора, у овом раду примењен је и *XGBoost.*

С обзиром да напредне статистичке анализе, укључујући методе машинског учења, дају значајан увид у промену нивоа мора, узет је рад Ниевеса и сарадника [3] као релевантан за ово истраживање. Уз помоћ машинског учења, главни циљ је одредити склоност ка повећању и смањењу нивоа мора у годинама које предстоје. Примењен је модел регресије Гаусовог процеса (GP) и рекурентна неуронска мрежа са јединицама дуготрајне меморије. Вршена је анализа утицаја отвореног океана на обалу, уз промену просечне температуре до 700 метара дубине. Аутори су предложили да би скуп података требало проширити и другим параметрима, као што је промена дебљине леда, салинитет и утицај глобалног загревања на површинске температуре. Сходно томе, наведени предлози су узети у обзир и у овом раду.

Рад [4] разматра примену техника машинског учења на проблеме регресије који се најчешће уочавају при анализи временских серија података. Регионалне варијације у порасту нивоа мора су знатне и требале би се узети у разматрање при планирању будућег пораста нивоа мора. Из тог разлога, примењени су историјски подаци о нивоу мора са мерача плиме и осеке који се налазе на многобројним местима распрострањеним дуж шведске обале. Методе машинског учења које су коришћење представљају три различите вештачке неуронске мреже и вишеструку линеарну регресију и оне су универзалне, те се могу лако применити и на податке са других локација. Доказано је да су свеукупне перформансе ових алгоритама машинског учења задовољавајуће, често и боље од перформанси много скупљих нумеричких модела океана.

3. МЕТОД

У наредним поглављима описан је начин добијања коначног скупа података, начин на који су модели креирани и експерименти над испробаним моделима за предикцију.

3.1. Скуп података

С обзиром да су фактори који утичу на глобално загревање разнолики, било је потребно прикупити податке из различитих извора. У циљу добијања прецизнијих резултата, сваки од прикупљених скупова података најпре је детаљно анализиран и процесиран. Услед великог броја података, уочено је да постоји значајан број оних који су невалидни, непотпуни или захтевају додатну обраду.

Након завршеног претпроцесирања, спајањем претходно добијених скупова података и издвајањем релевантних атрибута, формиран је коначан скуп. Подаци су груписани на основу датума мерења на месечном нивоу, док GMSL представља циљно обележје на основу кога је формирана класа - *IsGMSLIncreased*. Поред ових података, као релевантни атрибути издвојили су се и:

- *Extent* - укупна површина морског леда изражена у јединици 10^6 km^2,
- *WaterTemp* - температура воде у °C,
- O2ml - засићеност воде кисеоником,
- SIO3 - концентрација силиката,
- NO3 - концентрација нитрата,
- *LandAvgTemp* - просечна температура,
- *LandAndOceanAvgTemp* - просечна температура копна и мора за дан мерења и
- CO2 - просечна концентрација угљен-диоксида на месечном нивоу.

3.2. Генерисање модела за предикцију утицаја глобалног загревања на пораста нивоа мора

Циљ овог рада огледа се у проналажењу модела који дају најбоље резултате за дати проблем. Ради њиховог проналажења, за предвиђање су коришћени сви алгоритми споменути у релевантној литератури.

Прво су испробани Наивни Бајес и метод потпорних вектора, а затим метод случајне шуме и *Bagging* са и

без оптимизације хипер-параметара. Највише времена посвећено је *XGBoost* моделу и његовој оптимизацији.

Као улаз у систем коришћен је претходно претпроцесиран скуп података. С друге стране, излаз из система чинили су обучени модели на основу којих се закључује да ли се десило повећање нивоа мора.

Код обучавања модела временских серија, улаз у систем састојао се такође од претпроцесираних података коначног скупа. За разлику од претходно обучених модела, излаз из овог система представља обучен модел који пружа информацију о тачном повећању средњег глобалног нивоа мора - GMSL. Модел врши предвиђање будућих тачака података на основу историјских података, а као временски корак узет је интервал од месец дана.

3.3. Подешавање хипер-параметара модела машинског учења

Ради постизања најбољих перформанси модела над коначним скупом података, било је неопходно прецизно подесити вредности хипер-параметара. Процес подешавања хипер-параметара укључује испробавање њихових различитих вредности, обуку модела за сваки од њих и избор најбољих вредности.

Најпре, испробан је Наивни Бајесов модел. С обзиром да је релативно једноставан класификациони алгоритам, он нема велик број хипер-параметара у поређењу са сложенијим алгоритмима, па самим тим за овај модел није вршено додатно подешавање.

Затим, за метод потпорних вектора подешен је само тип функције језгра - RBF. Она је корисна у ситуацијама када је однос између карактеристика и ознака класа сложен и не може се ефикасно одвојити линеарном хипер-равном у оригиналном простору обележја. Самим тим, њеним избором омогућене су флексибилније и нелинеарне границе одлучивања.

Следећи испробан алгоритам био је метод случајне шуме - *RandomForestClassifier*. Прво је тестиран без подешавања хипер-параметара, а затим су испробане њихове различите комбинације како би открило који од њих дају најбоље резултате. Ради проналажења оптималних вредности хипер-параметара, коришћен је *RandomizedSearchCV*. Модел је пружао најбоље вредности када је *n_estimators* подешен на 100, *max_features* на *sqrt* и *max_depth* на 10.

Након тога, испробан је и *BaggingClassifier*. Оптимизација хипер-параметара вршена је истраживањем различитих вредности и упоређивањем њихових резултата. Прво је као базни естиматор дефинисан *DecisionTreeClassifier*, а у случају када је *n_estimators* подешен на 95, *max_samples* на 0.9 и *random_state* на 9, *BaggingClassifier*-а је давао најбоље резултате.

На крају, испробан је и *XGBoost*. За проналажење оптималних вредности параметара коришћен је *RandomizedSearchCV*. За естиматор је подешен *XGBRegressor*, док је сваком од хипер-параметара додељен низ могућих вредности. Оптимални хипер-параметри уз које *XGBoost* даје најбоље предикције постигнути су подешавањем: *max_depth* на 3, *min_child_weight* на 6, *learning_rate* на 0.01, *eta* на 2, *subsample* на 1, *colsample_bytree* на 0.8, *objective* на *binary:logistic* и *eval_metric* на *mae*.

Додатно, како би се предвидео тачан пораст глобалне средње вредности нивоа мора, испробан је модел временских серија.

За предвиђање је коришћен *ForecasterAutoreg*, при чему је за регресор одабран *XGBRegressor*. Емпиријским путем изведен је закључак да предвиђање са кораком 8 даје најбоље резултате. У ситуацији када је *n_estimators* подешен на 500, *max_depth* на 10, *learning_rate* на 0.1 и *lags* на низ вредности од 1 до 10, модел је пружао најпрецизније предикције.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

Наивни Бајес и метод потпорних вектора били су први тренирани модели. Њихова евалуација показала је идентичан учинак тачности, а он је износио 54%. Резултати су приказани у табели 4.1, где 0 показује да се пораст нивоа мора није десио, а 1 да јесте.

За оба наведена случаја, израчуната је прецизност, одзив и *F* мера. Из добијених резултата може се уочити да, услед глобалног раста нивоа мора, модели боље предвиђају повећање нивоа мора у односу на претходни месец него опадање.

Резултати	**Прецизност**	**Одзив**	***F* мера**
0	0.47	0.41	0.4
1	0.60	0.65	0.6

Табела 4.1. *Резултати евалуације за Наивног Бајеса и SVM.*

У *XGBoost* је следећи тренирани модел чија је основна средња апсолутна грешка износила 50%. Грешка је смањена за 2% при првој итерацији модела и то без оптимизације параметара, али добијени резултати су прилично лошији у односу на претходна два модела.

Оптимизација хипер-параметара модела, допринела је побољшању предикције пораста нивоа мора и то за 10%. Ови резултати представљени су у табели 4.2.

Резултати	**Прецизност**	**Одзив**	***F* мера**
0	0.51	0.53	0.52
1	0.64	0.61	0.62

Табела 4.2. *Резултати евалуације за XGBoost са оптимизацијом хипер-параметара.*

Уз помоћ *XGBoost*-а, закључено је да највећи утицај на пораст нивоа мора имају кисеоник, који је саставни део воде, нитрата (NO3), као и температура копна.

Метод случајне шуме без оптимизације хипер-параметара није допринео бољим резултатима. Међутим, након њиховог подешавања, добијено је најбоље предвиђање за ситуацију када се посматра смањење нивоа мора и то је приказано у табели 4.3.

Резултати	Прецизност	Одзив	*F* мера
0	0.60	0.53	0.57
1	0.62	0.68	0.65

Табела 4.3. *Резултати евалуације за метод случајне шуме са оптимизацијом хипер-параметара.*

Последњи модел за који су разматрани резултати предикције је *Bagging*. Метод случајне шуме и овај метод имали су идентичне резултате без оптимизације хипер-параметара, док су резултати добијени након подешавања приказани су у табели 4.4.

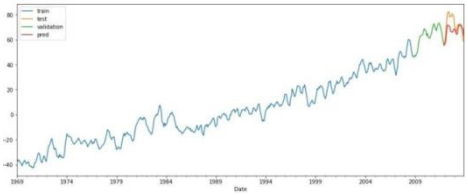
Резултати	Прецизност	Одзив	*F* мера
0	0.62	0.49	0.55
1	0.61	0.72	0.67

Табела 4.4. *Резултати евалуације за Bagging са оптимизацијом хипер-параметара.*

Након проучених, углавном лоших резултата свих обучених модела, изведен је закључак да сваки од њих даје приближно сличне резултате без обзира на подешавање хипер-параметара. Уз помоћ *XGBoost*-а, показано је да најзначајнији атрибути подударају са кључним факторима глобалног загревање.

Узимајући у обзир да претходни модели нису дали најбоље резултате, употребљене су временске серије које су допринеле бољим резултатима и анализама.

Фактори који највише утичу на коначне резултате су: временски кораци, дебљина леда, температура и молекули кисеоника у води. Ове вредности се скоро поклапају са оним добијеним од стране *XGBoost* модела. На слици 2, дат је приказ оригиналних и предвиђених вредности. Примећено је да модел лакше предвиђа опадање у односу на пораст нивоа мора.



Слика 2. *Предикција временских серија приказана црвеном линијом.*

Након анализе добијених резултата, уочено је да се нагли скок средње вредности нивоа мора догодио у години која је у односу на претходне имала рекордно високе температуре. У том периоду, модел је правио највише грешака, али тиме је уједно и потврђено да је температура најважнији фактор пораста GMSL-а. Међутим, за предикцију смањења нивоа мора, овај модел се показао као врло успешан.

5. ЗАКЉУЧАК

Услед све већег глобалног загревања и интересовања људи за ту тему, овај рад се бавио предвиђањем пораста нивоа мора.

Прикупљени су релевантни подаци из различитих извора где је сваки од скупова података имао заједнички атрибут који представља датум мерења на месечном нивоу. Извршено је њихово претпроцесирање и од таквих података формиран је коначан скуп. Први приступ се базирао на предикцији да ли се десио пораст нивоа мора у односу на претходни месец и који фактори су томе највише допринели. Модели машинског учења употребљени у овом приступу су: Наивни Бајес, метод потпорних вектора, метод случајне шуме, *Bagging* и *XGBoost*. Над коначним скупом података, упоређене су перформансе сваког од ових модела, док је за евалуацију тачности изабрана *F* мера. Примећено је да је након оптимизације хипер-параметара сваки модел давао приближне резултате, док се као најбољи издвојио *XGBoost*. Други приступ предвиђа тачне вредности средњег глобалног нивоа мора употребом временских серија. Оне су поред осталих фактора који утичу на предикцију, користиле и временски корак. За разлику од првог приступа, коришћење временских серија се испоставило као прецизније и корисније за даље унапређивање у циљу добијања бољих решења.

Један од недостатака овог рада лежи у ограниченом броју интегрисаних скупова података који имају различите нивое временске агрегације и не садрже константно мерене вредности. Уколико би се континуирано спроводио тренд акумулације података, претпоставка је да би модели били способнији да остваре прецизније прогнозе.

У циљу добијања бољих резултата, овај рад може бити проширен додатним факторима који утичу на прецизност модела. Такође, требало би посветити више времена анализи фактора за које модел предвиђа да највише утичу на пораст нивоа мора. Додатно, претпоставка је да би се поновним обучавање модела са подацима прикупљеним додатним мерењима и коришћењем временских серија са дневним кораком предвиђања, добили прецизнији резултати.

6. ЛИТЕРАТУРА

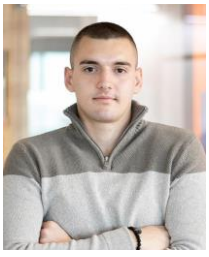
[1] Al. Balogun, N. Adebisi, “Geomatics, Natural Hazards and Risk”, vol. 12, 2021.

[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.

Кратка биографија:



Стефан Савић рођен је 1998. год. у Београду. Основне академске студије завршио је 2021. год. на Факултету техничких наука, на ком брани и мастер рад 2023. године из области Електротехнике и рачунарства.

Контакт: stefan.savic98@gmail.com