Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 528.44
DOI: https://doi.org/10.24867/08KG02Beker

UPOTREBA I VIZUELIZACIJA VELIKIH PODATAKA OTVORENOG TIPA ZA ANALIZU POGODNOSTI STANIŠTA EVROPSKE BUKVE NA PODRUČJU SRBIJE

USE AND VIZUALIZATION OF BIG OPEN SOURCE DATA FOR ANALYSIS OF HABITAT SUITABILITY OF EUROPEAN BEECH IN SERBIA

Teo Beker, *Fakultet tehničkih nauka, Novi Sad*

Oblast – GEODEZIJA I GEOMATIKA

Kratak sadržaj – *Studije pogodnosti staništa dobijaju povećani značaj usled ubrzanog menjanja zivotne sredine i globalnog zagrevanja. U ovom radu je istrenirano 6 modela mašinskog učenja na području cele Evrope i analizirani su i upoređeni rezultati na području Srbije. Finalni rezultati su vizualizovani.*

Ključne reči: *Neuronske mreže, pogodnost staništa, mašinsko učenje, veliki podaci, Fagus sylvatica*

Abstract – *Habitat suitability studies are becoming ever more valuable because of faster changing of environment and global warming. In this paper six machine learning models are trained on area of whole Europe and results are analyzed and compared on the area of Serbia. Final results are visualized.*

Keywords: *Neural networks; Habitat suitability; Machine learning; Big Data; Fagus sylvatica*

1. UVOD

Studija pogodnosti staništa [1] je jedna od osnovnih analiza kojima se bavi bioinformatika. Ona utvrđuje sličnosti između područja na kojima su zabeleženi primerci odredjene vrste sa područjima na kojima nisu ili se ne zna da li postoje, i na taj način tvrdi da sredine sa veoma sličnim uslovima bi trebale da budu pogodne za analiziranu vrstu. Rezultati su najčešće predstavljeni rasterskim kartama. Koriste se da se odrede granice regiona pogodnih za vrstu, za pronalaženje zona sa sličnim uslovima, za testiranje naučnih hipoteza i drugo.

Podaci otvorenog tipa igraju veliki značaj u moderno doba interneta i jakih kompjutera. Danas svako ko ima pristup podacima može da uzvede korisne analize, ili raznovrsne testove na dostupnim podacima. Besplatni podaci imaju veliki značaj u razvoju kako nauke tako i industrije. Veliki podaci sa druge strane nude velik materijal za objektivnije i informisanije analize [2].

Veliki podaci su u teoriji definisani sa 4V: Velocity (brzina), Variety (raznovrsnost), Volume (zapremina, veličina), Veracity (tačnosti i istinitost). Oni prave posebne probleme pri obradi i potrebno je imati mnogo kompjuterskih resursa i pažljivo pisati kod kada se radi sa njima.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Dušan Jovanović, docent.

Mašinsko učenje i neuronske mreže su vid veštačke inteligencije koji se koriste za kreiranje modela koji rešavaju probleme regresije, klasifikacije, klasterovanja i drugih… Postoje razni algoritmi u upotrebi, i po teoremi „Nema besplatnog ručka“ [3], ni jedan od njih ne može da reši svaki problem, i sam po sebi nije bolji od drugih.

U ovom radu su korišteni podaci otvorenog tipa da bi se upotrebom šest različitih modela mašinskog učenja dobili rezultati studije pogodnosti staništa za Evropsku bukvu. Analize su urađene na Evropskom nivou i na nivou Srbije, i na kraju su upoređeni.

2. PODACI

Za studiju korišteni su podaci otvorenog tipa, dostupni za celu Zemlju. Modeli su trenirani na nivou cele Evrope da bi bio uzet u obzir ceo spektar staništa koje Evropska bukva obuhvata.

2.1. Zemljište i teren

Za podatke o terenu korišteni su besplatno dostupni podaci Japanske agencije za istraživanje svemira (JAXA) [4]. Podaci su među najpreciznijim besplatno dostupnim podacima, u prostornoj rezoluciji od 30m, i visinskoj preciznosti od oko 5m. Podaci pokrivaju celu Zemlju i ceo skup podataka iznosi 1.6 TB.

Za zemljište je korišten SoilGRIDS skup podataka koji nudi podatke na globalnom nuvou u rezoluciji od 1km i 250m. Sastoji se iz oko 360 slojeva koji predstavljaju različite parametre i klasifikacije zemljišta. U radu se koristio skup sa parameterima zemljišta i USDA klasifikacija tipova zemljišta, jer se sastoji iz manje, preciznije određenih klasa zemljišta u poređenju sa WRB klasifikacijom.

2.2. Klima

Klimatski podaci su preuzeti sa WorldClim skupa podataka. Podaci su dostupni za ceo svet u rezoluciji od 30“, što iznosi oko 900m na ekvatoru. Postoje i futuristički klimatski scenariji u skupu, ali ova studija se fokusirala na trenutno stanje. Za potrebe predstavljanja Evropske bukve korišteno je 19 biovarijabli.

2.3. Upotreba zemljišta

Podaci o upotrebi zemljišta su dobijeni iz dva izvora. Područje pod šumama u Evropi je dobijeno sa Copernicus programa iz periodične Forest Type analize bazirane na satelitskim snimcima sa Sentinel misije. Podaci su

dostupni u rezoluciji od 30m, i šumu dele na zimzelenu, listopadnu i drugo.

Područje pod Evropskom bukvom je dobijeno iz FISE studije koja je na nivou Evrope uradila i predstavila analizu rasprostranjenosti oko 40 najbitnijih evropskih šumskih vrsta. Dostupna je u rezoluciji od oko 1km.

3. METODOLOGIJA

Metodologija je razvijena da svi podaci budu pripremljeni na pogodan način, da budu uzeti u obzir različiti aspekti modela dobijeni mašinskim učenjem i da svim modelima budu optimalno odabrani parametri i ocenjeni po najbitnijim metrikama.

3.1 Preprocesiranje podataka

Svi podaci su morali biti sastavljeni, preklopljeni da pikseli savršeno odgovaraju jedni drugima, svedeni na isti koordinatni referentni sistem: EPSG:4326, prostornu rezoluciju od 250m, vrednosti su im skalirane na raspon od 0 do 1 i null vrednosti obavezno obeležene i kasnije odstranjene.

3.2 Kreiranje skupa podataka

Dva trening skupa su napravljena da bi se u obzir uzeo željeni nivo generalizacije i nemogućnost određivanja tačnog područja koje ne pogoduje Evropskoj bukvi, usled ljudskog uticaja.

Područja koja sigurno nisu pogodna bukvi iz abiotičkih ili biotičkih razloga se mogu definisati kao područja pod šumama gde nije prisutna bukva. Ukoliko se razmatraju područja koja nisu pod šumama, ljudski uticaj je velik, bukva može biti odstranjena na primer zbog obradivih površina, izgradivih površina itd. Da bi se testirala ova hipoteza i uticaj odabira trening skupa testirana su dva slučaja:

- **Grupa modela 1**: gde se odsustvo bukve definiše samo na području pod šumama, i drugi slučaj:
- **Grupa modela 2:** gde je za odsustvo uzet ceo region Evrope koji nije pod bukvom (sa smanjenim uzorkom da odnos uzoraka koji označavju prisustvo i odsustvo bukve budu isti).

3.3 Odabir algoritama

Analize su rađene na šest algoritama mašinskog učenja: logistička regresija, SVM, random forest [5], adaboost, gradient boosting i multilayer perceptron neuronska mreža [6].

Logistička regresija se može koristiti kao osnovni, najjednostavniji model za poređenje sa ostalim. SVM je primenjivan u prošlosti sa odličnim rezultatima, ali pošto se u ovoj studiji koriste veliki podaci, sa jako mnogo parametara, SVM može imati problema da se izbori sa različitim distribucijama vrednosti. Ansambl metodi adaboost, gradient boosting i random forest rešavaju probleme varijanse i pristrasnosti u podacima, i zahvaljući tome često daju odlične rezultate. Neuronske mreže se koriste kao rešenje gotovo svakog problema, i njihova fleksibilnost, opravdava kompleksnost treniranja.

3.4 Optimizacija parametara

Za optimizaciju je korišten grid search metod koji je na osnovu 6 i više vrednosti za svaki parametar tražio optimalan set parametara za svaki algoritam.

3.5 Ocena algoritama

Kako u ovoj primeni nije cilj napraviti model koji će pronaći lokaciju na kojoj se već nalazi bukva, nego treba i da zadovolji određen nivo generalizacije, tačnost (1) i F1 mera nisu jedini parametri koji se trebaju gledati. Preciznost (2) daje inverznu meru generalizacije dok specifičnost ili TNR (eng. *true negative rate*) (3) daje direktnu meru generalizacije. Sve mere se moraju sagledati da bi se dobio dobar model.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made} \quad (1)$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (2)$$

$$Specificity = \frac{TrueNegative}{FalsePositive + TrueNegative} \quad (3)$$

4. IZVEDBA

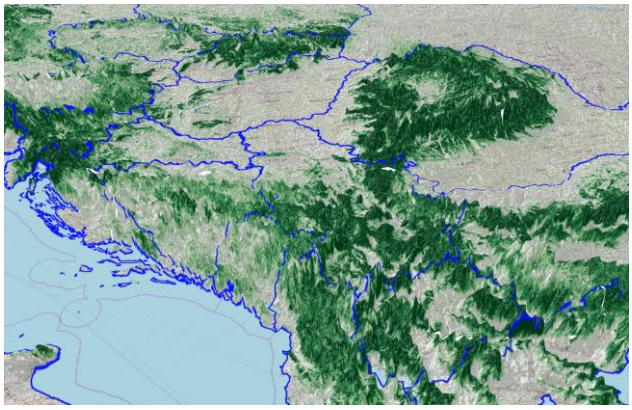
Studija je izvršena upotrebom python programskog jezika, BaseX sistema za upravljanje XML bazama podataka, GDAL [7] rešenjem za obradu prostornih podataka. Mašinsko učenje je primenjivano upotrebom scikit-learn, Tensor Flow i Keras softverskih bibilioteka, a preprocesiranje podataka je rađeno pomoću pandas [8] i numpy, kao i mnogih propratnih softverskih biblioteka.

Skup podataka je podeljen na dva dela, jedan za treniranje i jedan za testiranje. Svi modeli su trenirani koristeći kros-validaciju, i na kraju su testirani na test skupu podataka.

5. REZULTATI

Rezultati na testu su pokazali da je tačnost mnogo bolja na modelima iz prve grupe, obučavane na područjima samo pod šumama, a generalizacija na modelima iz druge. No ovaj test bez dodatnih informacija bi lako doveo do zablude. Zbog visoke prostorne korelacije, i smanjene varijanse pri odabiru područja samo pod šumama test set ne daje pravu sliku. Iz tog razloga urađene su analize koristeći podatke cele Evrope koji daju realističnije i logičnije rezultate.

Modeli iz grupe 2 su selektivniji, i više tačni. Za neuronsku mrežu u grupi 2 selektivnost je toliko visoka da se željena generalizacija nikako ne može postići ni spuštanjem praga selekcije. Random forest model je zadržao tu fleksibilnost, pa se kod njega oba modela mogu podesiti da zadovolje željene kriterijume



Slika 1. *3D prikaz rezultata modela random forest na području Srbije i Balkana (Intenzitet zelene boje prikazuje meru pogodnosti staništa za bukvu)*

generalizacije. SVM, adaboost i gradient boosting modeli nisu ostvarili svoj pun potencijal.
Da bi se iz njih dobili bolji rezultati neophodno je dublje modelovanje podataka i dodatna optimizacija parametera. Boosting metodi: Adaboost i gradient boosting algoritmi su rešenja koja su napravljena da umanje problem pristrasnosti podataka, koji se u ovoj studiji nije javio.
Random forest je ansambl metod koji umanjuje uticaj varijanse u podacima, koja je bila primarni problem u ovom zadatku. Zbog navedenih razloga boosting metodi se nisu pokazali dobro kao random forest.

Najbolji rezultati ostvareni u ovoj studiji su pomoću neuronske mreže, MLP, i zatim pomoću random forest modela. MLP konstantno kroz celu studiju ostvaruje najbolje rezultate po tačnosti, preciznosti i F1 meri. Dok je generalizacija (mala specifičnost) bitna na nivou Evrope, na području Srbije igra mnogo manju ulogu. Bitnije da rezultati budu tačni i precizni, jer već veliki deo Srbije je prekriven bukvom.
Poredeći rezultate dobijene na području Srbije (Tabela 1) i rezultate na nivou Evrope (Tabela 2), može se lako zapaziti pad u tačnosti, blago poboljšanje u preciznosti, drastično povećanu specifičnost i blago smanjnje AUC metrika.
Specifičnost koja daje inverznu meru generalizacije je mnogo pojačana u Srbiji, iz razloga što veći deo Srbije je veoma pogodan za bukvu, a severni deo u ravnici nije. To znači da ostaje jako malo prostora za generalizaciju, koja u ovom slučaju znači da se područja koja nisu pod šumom bukve klasifikuju kao pogodna za bukvu. Povećana preciznost takođe potvrđuje prethodni zaključak.
Na slici 2 se vidi poređenje rezultata dobijenih MLP modelom uz kalibraciju praga klasifikacije na 0.2 (umesto 0.5), i očekivanih i željenih rezultata dobijenih FISE studijom.
Rezultati su vrlo slični vizuelno i pokazuju da mašinsko učenje i veliki podaci imaju potencijal da dopune i automatizuju neke od koraka u biogeografskom modelovanju.
Rezultati random forest modela na području Srbije i Balkana su dati u 3D vizualizaciji na slici 1.

Algoritam	Accuracy [%]	Precision [%]	F1 [%]	Recall/Sensitivity [%]	Specificity [%]	AUC [%]
Modeli grupe 1 (trenirani na području pod šumom)						
Logistic Regression	71.346	43.704	59.559	93.469	**64.895**	85.482
Random Forests	73.206	45.682	62.494	**98.883**	65.719	81.846
Multilayer Perceptron	**77.269**	**49.825**	**66.256**	98.853	70.975	**95.108**
Modeli grupe 2 (trenirani na celom području)						
Logistic Regression	73.405	45.548	60.732	91.099	**68.246**	85.755
Random Forests	78.543	51.274	67.711	**99.657**	72.387	**96.877**
Multilayer Perceptron	**82.790**	**56.861**	**72.092**	98.466	78.219	96.280

Tabela 1. Rezultati testa na nivou Srbije modela iz grupe 1 i 2

Algorithm	Accuracy [%]	Precision [%]	F1 [%]	Recall/Sensitivity [%]	Specificity [%]	AUC [%]
Models trained on the forest training data set						
Logistic Regression	82.442	14.710	25.380	92.419	17.891	93.799
Random Forests	93.495	32.963	49.337	**98.035**	6.657	99.084
Multilayer Perceptron	**95.028**	**39.204**	**55.983**	97.868	**5.067**	**99.103**
Models trained on the whole training data set						
Logistic Regression	90.785	24.343	38.123	87.867	9.118	95.598
Random Forests	95.017	39.279	56.301	**99.355**	5.128	**99.443**
Multilayer Perceptron	**96.165**	**45.630**	**62.191**	97.623	**3.884**	99.209

Tabela 2. Rezultati dobijeni na nivou cele Evrope za grupu 1 i 2

Slika 2. *Sa leve Strane dobijeni rezultati MLP modela sa pragom spuštenim sa 0.5 na 0.2, a sa desne strane očekivani rezultati (FISE RPP studija, prag 0.05)*

6. ZAKLJUČAK

Rezultati pokazuju visoku tačnost i dobar nivo generalizacije na području cele Evrope. Modeli SVM nisu istrenirani na zadovoljavajući nivo na području cele Evrope, a adaboost i gradient boosting su postigli dobre rezultate, ali nikako za takmičenje sa MLP i random forest modelima. MLP i random forest su ostvarili najbolje rezultate u radu po statistici, a takođe kada se uporede rezultati uz podešavanje praga klasifikacije, sa očekivanim dobijenim FISE studijom, vidi se visok nivo sličnosti i korelacije, što se može videti na slici 2.

Nakon analize urađene samo na području Srbije, razlika između random forest i MLP modela su uočljivije, dok logistička regresija ujednačeno zaostaje kao i na nivou cele Evrope. Rezultati ukazuju da napredni modeli kao što su MLP i random forest mogu sami da se izbore sa varijansom vrednosti i da ostvare tačnije rezultate na celom skupu podataka. Ipak MLP iz grupe 2 iako tačniji nije bio u mogućnosti da zadovolji željeni nivo generalizacije na području cele Evrope, dok random forest je zadržao tu fleksibilnost.

U ovom radu je ukratko prikazan postupak upotrebe metoda mašinskog učenja u klasifikaciji Evropske bukve na nivou cele Evrope. Rezultati dobijeni na nivou cele Evrope su upoređeni sa rezultatima dobijenim samo na nivou Srbije. Rezultati su prezentovani u 3D prostoru.

7. LITERATURA

[1] Franklin, Janet. „*Mapping species distributions: spatial inference and prediction“*. Cambridge University Press, 2010.

[2] Teo Beker, Master Thesis: *“Big Data and machine learning for global evaluation of habitat suitability of European forest species”*, Milano, Politecnico di Milano, 2019.

[3] Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." *IEEE transactions on evolutionary computation* 1.1 (1997): 67-82.

[4] Takaku, Junichi, Takeo Tadono, and Ken Tsutsui. "GENERATION OF HIGH RESOLUTION GLOBAL DSM FROM ALOS PRISM." *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* 2.4 (2014).

[5] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[6] Bengio, Yoshua. "Practical recommendations for gradient-based training of deep architectures." *Neural networks: Tricks of the trade*. Springer, Berlin, Heidelberg, 2012. 437-478.

[7] GDAL/OGR contributors. "GDAL/OGR geospatial data abstraction software library." *Open Source Geospatial Foundation* (2018).

[8] McKinney, Wes. "pandas: a foundational Python library for data analysis and statistics." *Python for High Performance and Scientific Computing* 14 (2011).

Kratka biografija:

Teo Beker rođen je u Novom Sadu 1991. god. Master rad na Politehničkom univerzitetu u Milanu iz Geoinformatike – oblast Data Science odbranio je 2019. godine. Na fakultetu tehničkih nauka završava smer Geodezija i Geomatika – oblast Vizuelizacija geoprostornih podataka.
kontakt: teobeker@hotmail.com