Tema: Klasifikacija za detekciju okolišnih rizika

Kolegij: Strojno učenje za okolišne primjene

Ishodi učenja

Nakon završetka ove cjeline studenti će moći:

- Primijeniti metode nadziranog učenja za klasifikaciju okolišnih rizika i opasnosti.
- Evaluirati klasifikacijske modele primjenom metrika preciznosti, odziva, F1-mjere i AUC-a.
- Implementirati i interpretirati algoritme kao što su logistička regresija, metoda potpornih vektora (SVM), slučajna šuma (Random Forests) i konvolucijske neuralne mreže (CNN).
- Upotrijebiti izlaze modela za potporu odlukama u okolišnom monitoringu i upravljanju rizicima.

Uvod

Klasifikacijski modeli središnji su alat u znanosti o okolišnim podacima jer omogućuju preslikavanje raznolikih opažanja u diskretne kategorije rizika, kao što su kontaminirano u odnosu na nekontaminirano, štetno cvjetanje algi u odnosu na normalne uvjete ili sigurna u odnosu na opasna stanja. Za razliku od regresije, koja predviđa kontinuirane ishode (npr. koncentracije onečišćujućih tvari), klasifikacija daje kategorijske oznake koje se mogu izravno upotrijebiti za procjenu rizika.

Okolišni skupovi podataka obično integriraju više izvora informacija, uključujući:

- fizikalno-kemijske parametre (npr. pH, otopljeni kisik, opterećenja hranjivim tvarima),
- meteorološke pokretače (npr. temperatura, vjetar, oborine),
- deskriptore korištenja zemljišta i topografije,
- indekse daljinskih istraživanja i hiperspektralna mjerenja.

Primjeri klasifikacijskih zadataka u ovom kontekstu uključuju utvrđivanje je li uzorak vode kontaminiran ili čist, kategorizaciju kvalitete zraka u regulatorne razrede, razlikovanje tipova zemljišnog pokrova iz satelitskih snimaka te označavanje očitanja senzora kao normalnih ili anomalnih.

Ti su problemi zahtjevni jer često uključuju visokodimenzionalne i korelirane prediktore, vremenske ili prostorne ovisnosti koje narušavaju pretpostavke neovisnosti te izraženu

neuravnoteženost klasa, budući da se opasni događaji obično pojavljuju rjeđe od sigurnih uvjeta. Za rješavanje tih izazova potrebne su robusna predobrada (imputacija nedostajućih vrijednosti, normalizacija, kodiranje kategorijskih varijabli) i odgovarajuće validacijske strategije (npr. prostorna ili vremenska blokovna unakrsna validacija) kako bi se spriječilo curenje informacija i dobile realistične procjene sposobnosti generalizacije [4].

Osim prediktivne točnosti, ključne su interpretabilnost i kvantifikacija nesigurnosti. Analize važnosti značajki osiguravaju da se modeli oslanjaju na ekološki smislene pokretače (npr. opterećenja hranjivim tvarima, temperaturne anomalije), dok raspršenost ansambla i intervali predviđanja daju mjere nesigurnosti koje podupiru donošenje odluka svjesno rizika. Naposljetku, ponavljajući pomaci domene (zbog sezonskih ciklusa, nadogradnji senzora ili novih događaja) zahtijevaju periodično ponovno treniranje, transferno učenje i kontinuiranu evaluaciju kako bi se tijekom vremena održala pouzdana izvedba [4, 5, 7].

Vrste algoritama

U okolišnoj analitici često se primjenjuje nekoliko porodica klasifikacijskih algoritama:

- Logistička regresija: linearni probabilistički model koji daje interpretabilne koeficijente za binarne ili višeklasne ishode.
- Support Vector Machines (SVMs): klasifikatori utemeljeni na margini koji koriste jezgre za razdvajanje klasa u linearnim ili nelinearnim prostorima.
- Decision Trees: hijerarhijski modeli koji dijele podatke u klase primjenom uzastopnih pragova temeljenih na značajkama.
- Random Forests and Gradient Boosting: ansambli stabala odlučivanja koji povećavaju prediktivnu snagu i smanjuju varijancu.
- Convolutional Neural Networks (CNNs): modeli dubokog učenja koji automatski izdvajaju hijerarhijske značajke iz slikovnih ili spektralnih podataka.

Evaluacijske metrike

Evaluacija klasifikacijskih modela u okolišnoj analitici zahtijeva više od ukupne točnosti, jer su ekološke opasnosti često rijetke, a skupovi podataka izrazito neuravnoteženi.

Točnost

- Udio ispravno klasificiranih uzoraka među svim uzorcima.
- Korisna je kada su klase uravnotežene.

- Ograničenje: može zavarati kod neuravnoteženih podataka (npr. 95 % sigurnih u odnosu na 5 % opasnih: trivijalni klasifikator koji uvijek predviđa „sigurno" već postiže 95 % točnosti).

Preciznost

- Udio predviđenih pozitivnih primjera koji su stvarno pozitivni.
- Visoka preciznost znači malo lažnih alarma.
- Važna je kada su lažno pozitivni ishodi skupi (npr. nepotrebni zahvati sanacije).

Odziv (osjetljivost)

- Udio stvarnih pozitivnih primjera koji su ispravno identificirani.
- Visok odziv znači malo propuštenih opasnosti.
- Ključan je u okolišnom monitoringu, gdje lažno negativni ishodi (npr. propuštena izlijevanja nafte, neotkrivena kontaminacija) mogu imati ozbiljne posljedice.

F1-mjera

- Harmonijska sredina preciznosti i odziva.
- Uravnotežuje kompromis između otkrivanja opasnosti i izbjegavanja lažnih alarma.
- Široko se upotrebljava u neuravnoteženim skupovima podataka za sažimanje izvedbe.

ROC–AUC (Receiver Operating Characteristic)

- Mjeri kompromis između stope stvarno pozitivnih i stope lažno pozitivnih primjera pri različitim pragovima.
- Daje globalni prikaz izvedbe klasifikatora.
- Ograničenje: može biti pretjerano optimističan u izrazito neuravnoteženim postavkama.

PR–AUC (Precision–Recall)

- Obuhvaća odnos između preciznosti i odziva pri različitim pragovima.
- Informativniji je od ROC–AUC-a kada su anomalije rijetke.
- Osobito je važan u detekciji ekološkog rizika, gdje su opasni događaji rijetki, ali velikog učinka.

Okolišni problemi

Mnogi ključni okolišni izazovi mogu se formulirati kao klasifikacijski problemi, pri čemu različiti algoritmi nude komplementarne prednosti [4, 5].

Vodeni sustavi

Zadaci: detekcija štetnog cvjetanja algi, kategorizacija kontaminacije, klasifikacija trofičkog stanja.

Prikladni algoritmi:

- Logistička regresija: interpretabilna za kemijske/fizikalne prediktore [2];
- Ograničenje: slaba kod nelinearnih učinaka.
- Random Forests/Gradient Boosting: obuhvaćaju složene interakcije u miješanim hidro-kemijskim skupovima podataka [4];

Ograničenje: smanjena interpretabilnost.

Atmosferski sustavi

Zadaci: kategorizacija kvalitete zraka, detekcija perjanica industrijskog onečišćenja [4].

Prikladni algoritmi:

- SVMs: učinkoviti za visokodimenzionalne senzorske podatke [4];
  Ograničenje: računalni trošak pri velikom opsegu.
- Gradient Boosting: snažna prediktivna izvedba s neuravnoteženim podacima o kvaliteti zraka [5];

  Ograničenje: zahtijeva pažljivo podešavanje.

Šumarstvo i korištenje zemljišta

Zadaci: kartiranje rizika od šumskih požara, klasifikacija stresa vegetacije, detekcija deforestacije [4, 5].

Prikladni algoritmi:

- Decision Trees: jednostavni i interpretabilni za vegetacijske pragove;

  Ograničenje: skloni prenaučenosti.
- Random Forests: robusna izvedba na heterogenim prediktorima korištenja zemljišta [5];

  Ograničenje: manje transparentna logika odlučivanja.
- CNNs: idealni za snimke daljinskih istraživanja vegetacije/zemljišnog pokrova;
  Ograničenje: zahtijevaju velike količine podataka.

Zdravlje ekosustava

Zadaci: detekcija invazivnih vrsta iz eDNA, klasifikacija kategorija rizika od bolesti divljih životinja [5, 7].

Prikladni algoritmi:

- SVMs: dobro prilagođeni visokodimenzionalnim genomskim ili eDNA podacima [4]; Ograničenje: manje interpretabilni.
- Logistička regresija: interpretabilni osnovni model za čimbenike rizika od bolesti [4]; Ograničenje: slabija izvedba kod složenih interakcija.

U tim domenama odabir algoritma uravnotežuje prediktivnu točnost, interpretabilnost i računalnu učinkovitost, pri čemu se ansambli i modeli dubokog učenja ističu na složenim, visokodimenzionalnim podacima [5], dok linearni modeli i modeli temeljeni na stablima pružaju transparentnost i robusnost u jednostavnijim zadacima [4].

Detekcija opasnih/anomalnih objekata

Okolišne anomalije često odgovaraju rijetkim, neočekivanim i prostorno lokaliziranim događajima koji predstavljaju znatne ekološke rizike. Primjeri uključuju nagle skokove koncentracije lebdećih čestica, iznenadna štetna cvjetanja algi, širenje invazivnih vrsta ili pojavu mikroplastike i teških metala u vodenim i kopnenim sustavima. Ključni je izazov ekstremna neuravnoteženost klasa: opasni događaji mogu činiti manje od 5 % opažanja, zbog čega ih je teško detektirati konvencionalnim klasifikatorima. Osim toga, anomalije su često specifične za lokaciju, pa se modeli trenirani u jednoj regiji možda neće dobro generalizirati na drugu bez pažljive validacije.

Popis metoda

Za detekciju anomalija u okolišnim podacima primjenjuje se nekoliko pristupa:

- Statistička detekcija stršećih vrijednosti: identificira ekstremne vrijednosti uz pretpostavke o distribuciji.
- Ponovno uzorkovanje i troškovno osjetljivo učenje: ponovno uravnotežuju distribucije klasa primjenom prekomjernog uzorkovanja, poduzorkovanja ili dodjeljivanjem većih kazni lažno negativnim ishodima.
- Ansamblno učenje: random forests, gradient boosting i složeni modeli kombiniraju različite modele kako bi poboljšali robusnost na šum.

- Duboka detekcija anomalija: konvolucijski ili autoenkoderski modeli obuhvaćaju nelinearne obrasce i ističu neuobičajene prostorne ili spektralne potpise.

Primjeri specifičnih metoda

Tri reprezentativne tehnike prikazuju kako različiti pristupi rješavaju detekciju anomalija u praksi:

- One-Class SVM: uči granicu odlučivanja koja obuhvaća „normalne" podatke, označavajući uzorke izvan tog područja kao anomalije.
  Prednosti: učinkovit u visokodimenzionalnim prostorima, robustan na umjereni šum.
  Ograničenja: osjetljiv na izbor jezgre i hiperparametre; računalno skup na velikim skupovima podataka.
  Primjene: detekcija epizoda abnormalne kvalitete zraka, identifikacija invazivnih vrsta iz eDNA uzoraka [6, 7].
- Random Forest: ansambl stabala odlučivanja koji obuhvaća složene interakcije značajki i ističe stršeće vrijednosti na temelju pouzdanosti klasifikacije.
  Prednosti: skalabilan, robustan na prenaučenost, sposoban obraditi heterogene ulaze.
  Ograničenja: manje interpretabilan od jednostavnih stabala; izvedba ovisi o podešavanju i uravnoteženosti uzorka.
  Primjene: detekcija kontaminiranih tala iz kemijskih analiza, identifikacija neuobičajenih profila kvalitete vode [3].
- Autoenkoderi: neuronske mreže trenirane za rekonstrukciju ulaza; anomalije se detektiraju kada je pogreška rekonstrukcije visoka.
  Prednosti: obuhvaćaju složene nelinearne ovisnosti, prikladni za kontinuirani monitoring.
  Ograničenja: zahtijevaju dovoljno „normalnih" podataka za treniranje, računalno su zahtjevni, manje interpretabilni.
  Primjene: detekcija anomalija u stvarnom vremenu u monitoringu kvalitete vode [1].

Studija slučaja: Analiza rizika distribucije vrsta primjenom GBIF podataka o pojavljivanju i WorldClim bioklimatskih prediktora

Cilj

Ova studija slučaja prikazuje reproducibilan tijek rada za detekciju okolišnog rizika na temelju modeliranja distribucije vrsta (SDM). Cilj je identificirati okolišne uvjete povezane s prisutnošću

ciljane vrste i izgraditi model strojnog učenja koji procjenjuje vjerojatnost pojavljivanja u području interesa. U okviru ekološkog rizika, područja visoke predviđene prikladnosti mogu se interpretirati kao prioritetne zone za monitoring i prevenciju, osobito za invazivne vrste ili vektore bolesti.

Tijek rada implementiran je u Pythonu koristeći samo standardne biblioteke za znanost o podacima i otvoreno dostupne izvore podataka.+

Korištene biblioteke

Upotrijebit ćemo minimalan skup široko korištenih Python biblioteka za dohvat podataka, manipulaciju podacima, uzorkovanje rastera, strojno učenje i vizualizaciju. Odabrani je skup alata namjerno jednostavan i temelji se na "klasičnim" paketima koji se često koriste u primijenjenoj znanosti o podacima. U nastavku su popis i svrha korištenih biblioteka:

- **requests - koristi se za pristup web API-jima (GBIF) i preuzimanje klimatskih slojeva (WorldClim).**
- **pandas i numpy - koriste se za čišćenje podataka, filtriranje i inženjering značajki.**
- **rasterio - koristi se za čitanje GeoTIFF rastera i izdvajanje okolišnih varijabli na geografskim koordinatama.**
- **scikit-learn - pruža klasifikacijske modele i evaluacijske metrike.**
- **matplotlib - koristi se za jednostavne grafičke prikaze i interpretaciju modela.**

Taj odabir podupire transparentnost i olakšava izvođenje tijeka rada u standardnim okruženjima kao što su Jupyter Notebook ili Google Colab.

Izvor podataka

U ovom praktičnom vodiču usredotočujemo se na Vespa velutina (azijski stršljen), invazivnu vrstu koja se proširila u nekoliko europskih zemalja i može ugroziti lokalnu bioraznolikost i pčelarstvo jer se hrani oprašivačima poput medonosnih pčela. Zbog toga je otkrivanje područja u kojima je vjerojatno da će se vrsta sljedeća pojaviti praktičan primjer detekcije okolišnog rizika. Cilj vodiča jest izgraditi jednostavan model strojnog učenja koji uči okolišne uvjete povezane s prisutnošću vrste, a zatim proizvodi ocjenu prikladnosti staništa ili rizika za nove lokacije. Konkretnije, skup podataka sastoji se od zapisa prisutnosti (koordinate zemljopisne širine/dužine i metapodaci opažanja) za Vespa velutina unutar odabrane regije i vremenskog prozora (2000.–2024.), preuzetih putem javnog GBIF API-ja. Za izgradnju ulaznih značajki svaka točka pojavljivanja vrste obogaćuje se klimatskim prediktorima izdvojenima iz bioklimatskih varijabli WorldClim v2.1, koje

su globalni rasterski slojevi što sažimaju dugoročne obrasce temperature i oborina, često korištene u modeliranju distribucije vrsta. Konačni skup podataka stoga povezuje svaku lokaciju sa skupom klimatskih pokazatelja (npr. godišnja srednja temperatura, sezonalnost oborina) i omogućuje binarni klasifikacijski zadatak temeljen na prisutnosti u odnosu na pseudo-odsutnost (pozadinske) točke uzorkovane u istoj regiji.

Izvori podataka: GBIF Occurrence API https://techdocs.gbif.org/en/openapi/v1/occurrence i WorldClim v2.1 Bioclim https://www.worldclim.org/data/worldclim21.html

Vrste podataka

Vodič kombinira dvije česte vrste geoprostornih podataka koje se često zajedno koriste u okolišnoj analitici: točkaste podatke o pojavljivanju (gdje je nešto opaženo) i rasterske podatke (prostorne slojeve koji opisuju okolišne uvjete posvuda u regiji).

- **Podaci o pojavljivanju** (točkasta opažanja): Prvi je skup podataka tablica GBIF zapisa pojavljivanja za Vespa velutina. U kodu je svako opažanje pohranjeno kao jedan redak s geografskim koordinatama (decimalLatitude, decimalLongitude) i osnovnim metapodacima: gbifID, year, basisOfRecord, datasetKey i (kada je dostupno) coordinateUncertaintyInMeters Te točke predstavljaju lokacije prisutnosti (oznaka 1), odnosno mjesta na kojima je vrsta stvarno zabilježena. Budući da GBIF uglavnom pruža podatke samo o prisutnosti, ti zapisi ne uključuju izravno potvrđene odsutnosti.

```
# ------------------------------
SPECIES_NAME = "Vespa velutina"  # invasive Asian hornet example
# Italy bounding box (lon_min, lat_min, lon_max, lat_max)
ROI = dict(lon_min=6.6, lat_min=36.6, lon_max=18.6, lat_max=47.2)

YEAR_FROM, YEAR_TO = 2000, 2024

taxon_key, match_json = gbif_taxon_key(SPECIES_NAME)

pres = fetch_gbif_occurrences(
    taxon_key, ROI, YEAR_FROM, YEAR_TO,
    max_points=6000
)

pres = pres.dropna(subset=["decimalLatitude", "decimalLongitude"]).copy()
pres = pres.drop_duplicates(subset=["decimalLatitude", "decimalLongitude"]).copy()

pres["label"] = 1
pres.head()
```

1 to 5 of 5 entries Filter

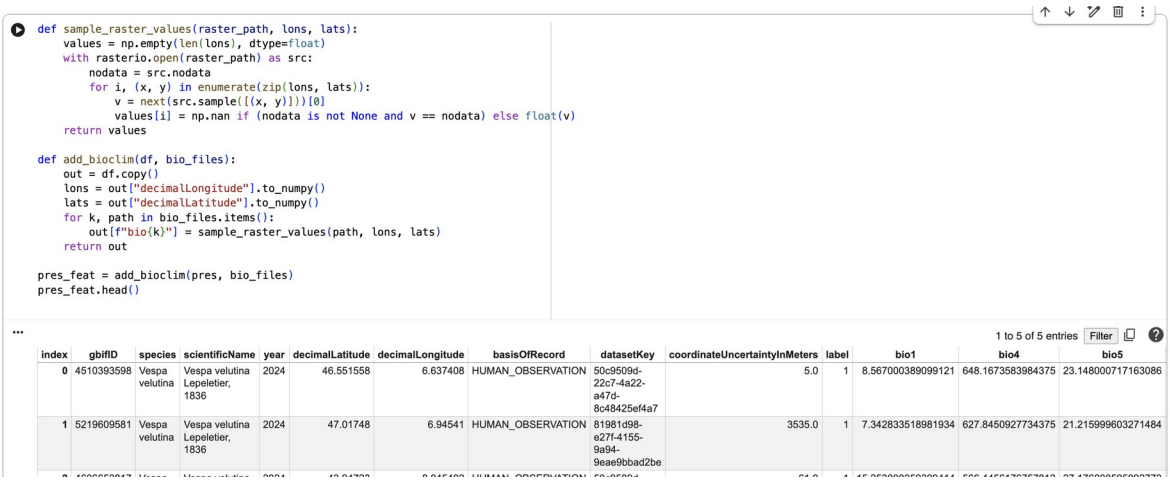
index	gbifID	species	scientificName	year	decimalLatitude	decimalLongitude	basisOfRecord	datasetKey	coordinateUncertaintyInMeters	label
0	4510393598	Vespa velutina	Vespa velutina Lepeletier, 1836	2024	46.551558	6.637408	HUMAN_OBSERVATION	50c9509d-22c7-4a22-a47d-8c48425ef4a7	5.0	1
1	5219609581	Vespa	Vespa velutina	2024	47.01748	6.94541	HUMAN_OBSERVATION	81981d98-e27f-4155-	3535.0	1

- **Rasterski podaci** (mrežasti okolišni slojevi): Drugi skup podataka sastoji se od WorldClim v2.1 bioklimatskih rastera pohranjenih kao GeoTIFF datoteke. Svaki raster je mreža koja pokriva regiju, pri čemu svaka ćelija sadrži klimatsku vrijednost. U ovom vodiču koristimo šest prediktora: BIO1, BIO4, BIO5, BIO12, BIO14, BIO15, koji će postati stupci bio1, bio4, bio5, bio12, bio14 i bio15. Rasterske vrijednosti izdvajaju se pomoću rasterio.

```
# WorldClim bioclim variables to use.
# https://www.worldclim.org/data/bioclim.html
# BIO1: Annual Mean Temperature
# BIO4: Temperature Seasonality (standard deviation ×100)
# BIO5: Max Temperature of Warmest Month
# BIO12: Annual Precipitation
# BIO14: Precipitation of Driest Month
# BIO15: Precipitation Seasonality (Coefficient of Variation)
BIO_VARS = [1, 4, 5, 12, 14, 15]

# We can download where we will be downloading our data
WC_URL = "https://geodata.ucdavis.edu/climate/worldclim/2_1/base/wc2.1_2.5m_bio.zip"

# ... download and unzip ...we need to keep only the files related to the BIO_VARS

bio_files = {}
for k in BIO_VARS:
    path = os.path.join(extract_dir, f"wc2.1_2.5m_bio_{k}.tif")
    bio_files[k] = path

bio_files
```

```
{1: 'data_sdm/wc2.1_2.5m_bio/wc2.1_2.5m_bio_1.tif',
 4: 'data_sdm/wc2.1_2.5m_bio/wc2.1_2.5m_bio_4.tif',
 5: 'data_sdm/wc2.1_2.5m_bio/wc2.1_2.5m_bio_5.tif',
 12: 'data_sdm/wc2.1_2.5m_bio/wc2.1_2.5m_bio_12.tif',
 14: 'data_sdm/wc2.1_2.5m_bio/wc2.1_2.5m_bio_14.tif',
 15: 'data_sdm/wc2.1_2.5m_bio/wc2.1_2.5m_bio_15.tif'}
```

- Kako se kombiniraju dvije vrste podataka (izdvajanje rastera u točkama): Korak integracije provodi se obogaćivanjem svake GBIF točke rasterskim vrijednostima: za svaku koordinatu (lon,lat)(lon, lat)(lon,lat) očitavamo vrijednost iz svakog WorldClim rastera i dodajemo je kao novi stupac značajke. Time nastaje standardna tablica za strojno učenje u kojoj je svaki redak jedna lokacija, a svaki prediktor klimatska varijabla.



```
def sample_raster_values(raster_path, lons, lats):
    values = np.empty(len(lons), dtype=float)
    with rasterio.open(raster_path) as src:
        nodata = src.nodata
        for i, (x, y) in enumerate(zip(lons, lats)):
            v = next(src.sample([(x, y)]))[0]
            values[i] = np.nan if (nodata is not None and v == nodata) else float(v)
    return values

def add_bioclim(df, bio_files):
    out = df.copy()
    lons = out["decimalLongitude"].to_numpy()
    lats = out["decimalLatitude"].to_numpy()
    for k, path in bio_files.items():
        out[f"bio{k}"] = sample_raster_values(path, lons, lats)
    return out

pres_feat = add_bioclim(pres, bio_files)
pres_feat.head()
```

1 to 5 of 5 entries Filter

index	gbifID	species	scientificName	year	decimalLatitude	decimalLongitude	basisOfRecord	datasetKey	coordinateUncertaintyInMeters	label	bio1	bio4	bio5
0	4510393598	Vespa velutina	Vespa velutina Lepeletier, 1836	2024	46.551558	6.637408	HUMAN_OBSERVATION	50c9509d-22c7-4a22-a47d-8c48425ef4a7	5.0	1	8.567000389099121	648.1673583984375	23.148000717163086
1	5219609581	Vespa velutina	Vespa velutina Lepeletier, 1836	2024	47.01748	6.94541	HUMAN_OBSERVATION	81981d98-e27f-4155-9a94-9eae9bbad2be	3535.0	1	7.342833518981934	627.8450927734375	21.215999603271484

Budući da GBIF uglavnom pruža zapise samo o prisutnosti, vodič stvara pozadinske točke u istoj regiji i dodjeljuje im oznaku 0. Te se točke koriste kao zamjena za lokacije neprisutnosti, čime se omogućuje binarni klasifikacijski skup podataka. Te se točke koriste kao zamjena za lokacije neprisutnosti, čime se omogućuje binarni klasifikacijski skup podataka.

```
bg = random_points_bbox(n_bg, ROI, seed=42)
bg["label"] = 0

bg_feat = add_bioclim(bg, bio_files)
bg_feat.head()
```

	decimalLongitude	decimalLatitude	label	bio1	bio4	bio5	bio12	bio14	bio15
0	15.887473	46.851860	0	9.194166	768.208618	24.972	842.0	30.0	38.745869
1	11.866541	43.871469	0	10.660334	650.370911	25.180	1430.0	64.0	30.592279
2	16.903175	37.555253	0	NaN	NaN	NaN	NaN	NaN	NaN
3	14.968416	43.947677	0	NaN	NaN	NaN	NaN	NaN	NaN
4	7.730128	42.998380	0	NaN	NaN	NaN	NaN	NaN	NaN

Čišćenje podataka

Skupovi podataka o pojavljivanju često zahtijevaju osnovni korak čišćenja jer mogu sadržavati duplicirane koordinate (ista lokacija prijavljena više puta), nesigurne koordinate niske prostorne preciznosti i, u rijetkim slučajevima, točke koje se nalaze izvan odabranog područja interesa. Budući da je cilj ovog vodiča povezati okolišne uvjete s prisutnošću vrste, netočni ili ponovljeni zapisi mogli bi pristraniti analizu i dobiveni model strojnog učenja. Zbog toga tijek rada primjenjuje standardne filtre uklanjanjem zapisa s nedostajućom zemljopisnom širinom ili dužinom, zadržavanjem samo opažanja unutar ROI graničnog okvira, uklanjanjem dupliciranih parova koordinata i, kada je dostupno, opcionalnim isključivanjem točaka s vrlo velikom koordinatnom nesigurnošću.

Konačna tablica za strojno učenje za treniranje

Točke prisutnosti i pozadinske točke (obje s izdvojenim BIO prediktorima) kombiniraju se u jedan skup podataka koji se koristi za treniranje. Matrica ulaznih značajki X sadrži odabrane bioklimatske varijable izdvojene iz WorldClim-a (BIO prediktore), dok ciljni vektor y sadrži binarne oznake koje označavaju prisutnost u odnosu na pseudo-odsutnost. Time nastaje standardni skup podataka za nadzirano učenje prikladan za treniranje klasifikacijskih modela.

```
# --------------------------------
# 7) Build final ML table (X,y)
# --------------------------------
feature_cols = [f"bio{k}" for k in BIO_VARS]
cols = ["label"] + feature_cols

data = pd.concat([pres_feat[cols], bg_feat[cols]], ignore_index=True)
data = data.dropna().copy()

X = data[feature_cols].to_numpy()
y = data["label"].to_numpy()

data["label"].value_counts()
```

```
...
```

label	count
0	3778
1	1566

dtype: int64

Podjela na skup za treniranje/testni skup i osnovni model (Logistička regresija)

Kako bi se evaluirala sposobnost modela da se generalizira na neviđene podatke, skup podataka dijeli se na skup za treniranje i testni skup. Koristi se stratificirana podjela kako bi se očuvao udio točaka prisutnosti i pseudo-odsutnosti u oba podskupa, čime se osigurava da procjene izvedbe nisu pristrane zbog neuravnoteženosti klasa. Logistička regresija koristi se kao prvi osnovni model jer je jednostavan i interpretabilan, daje probabilistička predviđanja i snažna je referentna metoda za probleme binarne klasifikacije. Evaluacijske metrike: Budući da su ekološki skupovi podataka o prisutnosti često neuravnoteženi, izvedba modela procjenjuje se metrikama koje su informativne i izvan same točnosti, uključujući ROC-AUC, preciznost i odziv, F1-mjeru te matricu zabune. Time se podupire smislenija interpretacija prediktivne izvedbe u kontekstu detekcije okolišnog rizika.

```
# ------------------------------
# 8) Train/test split + Logistic Regression
# ------------------------------
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.25, random_state=0, stratify=y
)

logreg = LogisticRegression(max_iter=3000, class_weight="balanced")
logreg.fit(X_train, y_train)

proba_lr = logreg.predict_proba(X_test)[:, 1]
pred_lr = (proba_lr >= 0.5).astype(int)

print("LogReg ROC-AUC:", roc_auc_score(y_test, proba_lr))
print(classification_report(y_test, pred_lr, digits=3))
print("Confusion matrix:\n", confusion_matrix(y_test, pred_lr))
```

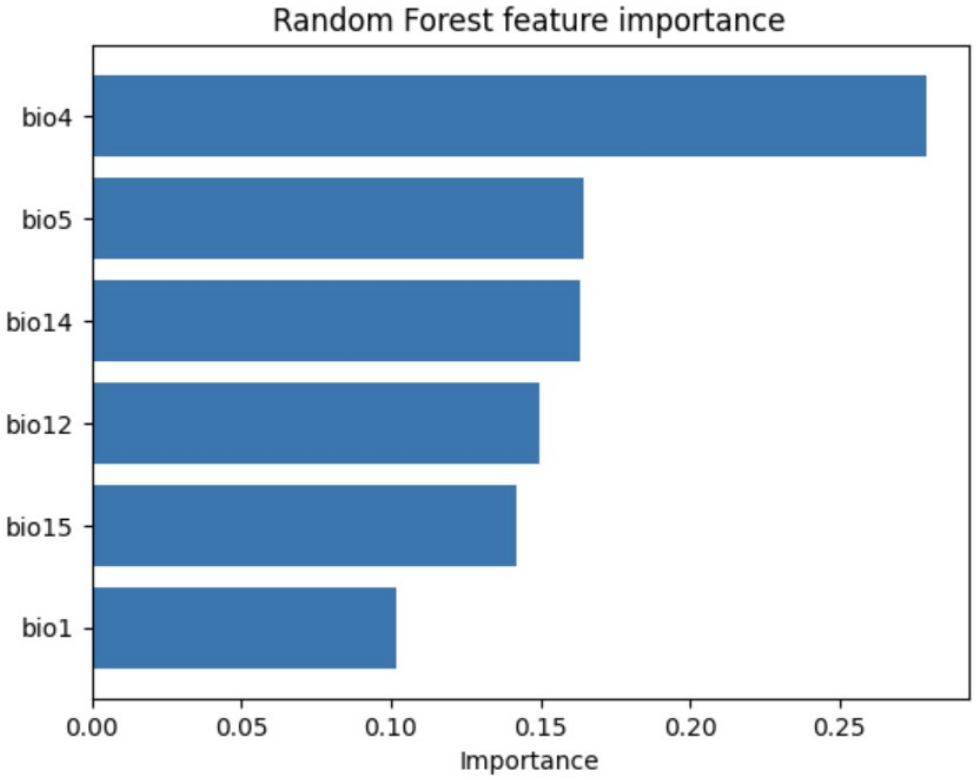
```
LogReg ROC-AUC: 0.9246512131422616
              precision    recall  f1-score   support

           0      0.952     0.820     0.881       945
           1      0.674     0.900     0.771       391

    accuracy                          0.844      1336
   macro avg      0.813     0.860     0.826      1336
weighted avg      0.871     0.844     0.849      1336

Confusion matrix:
 [[775 170]
 [ 39 352]]
```

Random Forest i važnost značajki

Random Forest uvodi se kao drugi model jer su odnosi između vrsta i okoliša često nelinearni i mogu uključivati interakcije među prediktorima. Random Forest obično dobro djeluje na tabličnim okolišnim skupovima podataka i robustan je na složene granice odlučivanja. Random Forest također pruža praktičan oblik interpretabilnosti putem važnosti značajki, koja ističe koji klimatski prediktori najviše doprinose razlikovanju točaka prisutnosti od točaka pseudo-odsutnosti. To podupire ekološku interpretaciju, primjerice identificiranjem povezanosti s toplijim uvjetima, osjetljivošću na oborine u sušnoj sezoni ili ulogom sezonskih ograničenja. U cjelini, ovaj korak učvršćuje perspektivu „objašnjive detekcije rizika".

```
# ------------------------------
# 9) Random Forest + feature importance
# ------------------------------
rf = RandomForestClassifier(
    n_estimators=400,
    random_state=0,
    class_weight="balanced_subsample",
    n_jobs=-1
)
rf.fit(X_train, y_train)

proba_rf = rf.predict_proba(X_test)[:, 1]
pred_rf = (proba_rf >= 0.5).astype(int)

print("RF ROC-AUC:", roc_auc_score(y_test, proba_rf))
print(classification_report(y_test, pred_rf, digits=3))
print("Confusion matrix:\n", confusion_matrix(y_test, pred_rf))

# Importance plot
importances = rf.feature_importances_
order = np.argsort(importances)

plt.figure()
plt.barh([feature_cols[i] for i in order], importances[order])
plt.title("Random Forest feature importance")
plt.xlabel("Importance")
plt.show()
```

```
RF ROC-AUC: 0.988926778440845
              precision    recall  f1-score   support

           0      0.970     0.972     0.971       945
           1      0.933     0.928     0.931       391

    accuracy                          0.960      1336
   macro avg      0.952     0.950     0.951      1336
weighted avg      0.960     0.960     0.960      1336

Confusion matrix:
 [[919  26]
 [ 28 363]]
```



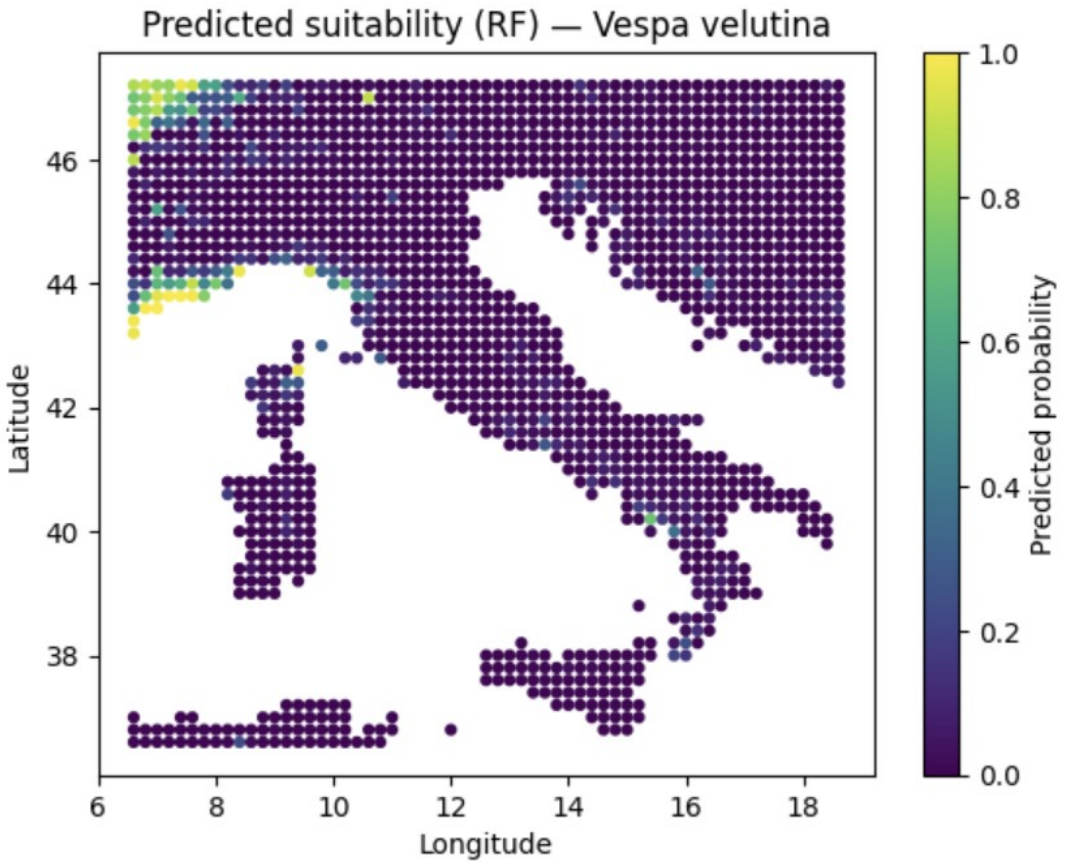
Model Random Forest postiže vrlo snažnu izvedbu na ovom skupu podataka, s ROC–AUC-om od 0,99 i ukupnom točnošću od 0,96, što upućuje na to da klimatski prediktori jasno razdvajaju točke prisutnosti od točaka pseudo-odsutnosti. Matrica zabune pokazuje samo ograničen broj pogrešaka (26 lažno pozitivnih i 28 lažno negativnih), a klasifikacijsko izvješće potvrđuje visoku preciznost i odziv za obje klase. Prema grafikonu važnosti značajki, BIO4 (sezonalnost temperature) najutjecajnija je varijabla, što sugerira da je varijabilnost temperature tijekom godine ključan čimbenik u razlikovanju prikladnih i neprikladnih uvjeta za Vespa velutina. Ostali relevantni prediktori uključuju BIO5 (maksimalna temperatura najtoplijeg mjeseca) i varijable povezane s

oborinama (BIO12, BIO14, BIO15), dok se BIO1 (godišnja srednja temperatura) sam po sebi čini manje informativnim. Ti rezultati podupiru ideju da prikladnost za Vespa velutina nije određena samo prosječnim klimatskim razinama, nego i sezonalnošću i ekstremima, što je u skladu s ekološkim očekivanjima za uspostavu invazivnih vrsta.

Prostorno predviđanje

Ključna prednost modeliranja distribucije vrsta jest to što se istrenirani model može primijeniti na neuzorkovane lokacije. Kako bi se to prikazalo, preko područja interesa generira se mreža točaka i ocjenjuje pomoću istreniranog klasifikatora, čime nastaje prostorna površina predviđanja. Predviđena vjerojatnost može se interpretirati kao ocjena prikladnosti staništa, zamjena za vjerojatnost uspostave ili pokazatelj prioriteta monitoringa. Visoke predviđene vrijednosti odgovaraju područjima u kojima bi vrsta mogla vjerojatno opstati, čak i ako ondje još nisu prijavljena pojavljivanja.

```
# ---------------------------------
# 10) Optional: quick suitability map on a coarse grid
#     (no geopandas; just a scatter heatmap)
# ---------------------------------
def make_grid(roi, step_deg=0.2):
    lons = np.arange(roi["lon_min"], roi["lon_max"] + 1e-9, step_deg)
    lats = np.arange(roi["lat_min"], roi["lat_max"] + 1e-9, step_deg)
    grid = [(x, y) for y in lats for x in lons]
    return pd.DataFrame(grid, columns=["decimalLongitude", "decimalLatitude"])

grid = make_grid(ROI, step_deg=0.2)
grid_feat = add_bioclim(grid, bio_files)
grid_feat = grid_feat.dropna(subset=feature_cols).copy()

Xg = grid_feat[feature_cols].to_numpy()
grid_feat["suitability"] = rf.predict_proba(Xg)[:, 1]

plt.figure()
plt.scatter(grid_feat["decimalLongitude"], grid_feat["decimalLatitude"], c=grid_feat["suitability"], s=12)
plt.title(f"Predicted suitability (RF) — {SPECIES_NAME}")
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.colorbar(label="Predicted probability")
plt.show()
```



Zadaci za vježbu

- Ponovite isti tijek rada iz ovog vodiča za drugu vrstu po vlastitom izboru (na primjer invazivnu vrstu, oprašivača ili vrstu relevantnu za vašu regiju).
- Zamijenite SPECIES_NAME u kodu, promijenite ROI za svoju regiju i zadržite isti vremenski prozor (2000.–2024.) te pokrenite cijeli pipeline: preuzmite GBIF zapise pojavljivanja (točke prisutnosti), očistite koordinate, generirajte točke pseudo-odsutnosti, izdvojite iste WorldClim prediktore (bio1, bio4, bio5, bio12, bio14, bio15) i istrenirajte i Logistička regresija i Random Forest. Za odabir prikladne vrste provjerite vraća li vaš GBIF upit dovoljno zapisa (najmanje ~200 točaka nakon čišćenja).
- U kratkom izvješću uključite: (i) broj korištenih točaka prisutnosti, (ii) ROC–AUC i matricu zabune za oba modela te (iii) grafikon važnosti značajki modela Random Forest.

- Naposljetku, odgovorite na ova pitanja: Koji model ima bolju izvedbu i zašto? Koja je BIO varijabla najvažnija za vašu vrstu? Sugeriraju li rezultati da je vrsta više uvjetovana ograničenjima povezanima s temperaturom ili oborinama?

Provjera znanja

Koji je glavni cilj primjene klasifikacijskih modela u detekciji okolišnog rizika?

A. Predvidjeti kontinuirane okolišne varijable kao što su temperatura ili koncentracija onečišćujućih tvari
B. Grupirati okolišna opažanja u neoznačene klastere
C. Dodijeliti opažanja unaprijed definiranim kategorijama rizika (npr. opasno u odnosu na sigurno)
D. Smanjiti dimenzionalnost okolišnih skupova podataka

Zašto je ukupna točnost često nedostatna za evaluaciju klasifikacijskih modela u okolišnim primjenama?

A. Zato što su okolišni podaci uvijek šumni
B. Zato što se točnost ne može izračunati za višeklasne probleme
C. Zato što su okolišni skupovi podataka često izrazito neuravnoteženi, s rijetkim opasnim događajima
D. Zato što točnost zanemaruje interpretabilnost modela

Koja je evaluacijska metrika najkritičnija kada je cilj minimizirati propuštene okolišne opasnosti (lažno negativne ishode)?

A. Preciznost
B. Točnost
C. Odziv (osjetljivost)
D. ROC–AUC

Koji je algoritam osobito prikladan za detekciju izlijevanja nafte ili morskog otpada na satelitskim ili dronskim snimkama?

A. Logistička regresija
B. Support Vector Machines

C. Random Forests
D. Convolutional Neural Networks (CNNs)

Što najbolje opisuje ulogu izlaza klasifikacijskih modela u upravljanju okolišnim rizikom?

A. U potpunosti zamjenjuju stručnu prosudbu
B. Daju točne prognoze budućih okolišnih uvjeta
C. Podupiru donošenje odluka identificiranjem i određivanjem prioriteta kategorija rizika
D. Uklanjaju nesigurnost u sustavima monitoringa

Dodatni izvori

Uvodni vodiči o prognoziranju u Pythonu (npr. scikit-learn, pandas).

Machine Learning - A First Course for Engineers and Scientists, Andreas Lindholm, Niklas Wahlström, Fredrik Lindsten, and Thomas B. Schön

Deep Learning with Python, Second Edition, François Chollet

Applications of Deep Neural Networks with Keras, Jeff Heaton

Understanding Machine Learning: From Theory to Algorithms, Shai Shalev-Shwartz and Shai Ben-David

Understanding Machine Learning: From Theory to Algorithms, Shai Shalev-Shwartz and Shai Ben-David

Bibliografija

[1] El-Shafeiy, E., Alsabaan, M., Ibrahem, M. I., & Elwahsh, H. (2023). Real-time anomaly detection for water quality sensor monitoring based on multivariate deep learning technique. Sensors, 23(20), 8613.
[2] Hu, X., Luo, H., Guo, M., & Wang, J. (2022). Ecological technology evaluation model and its application based on Logistic Regression. Ecological Indicators, 136, 108641.
[3] Madani, A., Hagage, M., & Elbeih, S. F. (2022). Random Forest and Logistic Regression algorithms for prediction of groundwater contamination using ammonia concentration. Arabian Journal of Geosciences, 15(20), 1619.

[4] Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing, 39 (9), 2784–2817.

[5] Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. Methods in Ecology and Evolution, 14 (4), 994–1016.

[6] Scholkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443–1471.

[7] Zinger, L., Bonin, A., Alsos, I.G., Ba´lint, M., Bik, H., Boyer, F., Chariton, A.A., Creer, S., Coissac, E., Deagle, B.E. & De Barba, M., (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. Molecular ecology, 28(8), pp.1857-1862.