Tema: Klasifikacija za zaznavanje okoljskih tveganj

Predmet: Strojno učenje za okoljske aplikacije

Učni izidi

Po zaključku te enote bodo študenti znali:

- Uporabiti metode nadzorovanega učenja za klasifikacijo okoljskih tveganj in nevarnosti.
- Ovrednotiti klasifikacijske modele z merami preciznosti, občutljivosti, F1 in AUC.
- Implementirati in interpretirati algoritme, kot so logistična regresija, SVM, naključni gozdovi in CNN.
- Uporabiti izhode modelov kot podporo pri okoljskem spremljanju in odločanju o obvladovanju tveganj.

Uvod

Klasifikacijski modeli so osrednje orodje okoljske podatkovne znanosti, saj omogočajo razvrščanje raznovrstnih opazovanj v diskretne kategorije tveganja, kot so onesnaženo oziroma neonesnaženo okolje, škodljivo cvetenje alg oziroma običajne razmere ali varna oziroma nevarna stanja. V nasprotju z regresijo, ki napoveduje zvezne izide (npr. koncentracije onesnaževal), daje klasifikacija kategorične oznake, ki jih je mogoče neposredno uporabiti pri ocenjevanju tveganja.

Okoljski podatkovni nabori navadno združujejo več virov informacij, med drugim:

- fizikalno-kemijske parametre (npr. pH, raztopljeni kisik, obremenitve s hranili),
- meteorološke dejavnike (npr. temperaturo, veter, padavine),
- opisne spremenljivke rabe tal in topografije,
- indekse daljinskega zaznavanja in hiperspektralne meritve.

Primeri klasifikacijskih nalog v tem kontekstu vključujejo ugotavljanje, ali je vzorec vode onesnažen ali čist, razvrščanje kakovosti zraka v predpisane razrede, razlikovanje vrst pokrovnosti tal na satelitskih posnetkih ter označevanje senzorskih odčitkov kot običajnih ali odstopajočih.

Ti problemi so zahtevni, ker pogosto vključujejo visokorazsežne in korelirane napovedne spremenljivke, časovne ali prostorske odvisnosti, ki kršijo predpostavko neodvisnosti, ter močno neuravnoteženost razredov, saj se nevarni dogodki praviloma pojavljajo redkeje kot varne

razmere. Za obravnavo teh izzivov so potrebne robustna predobdelava (dopolnjevanje manjkajočih vrednosti, normalizacija, kodiranje kategoričnih spremenljivk) in ustrezne strategije validacije (npr. prostorska ali časovna blokovna navzkrižna validacija), ki preprečujejo uhajanje informacij in omogočajo realistične ocene posplošitvene uspešnosti [4].

Poleg natančnosti napovedovanja sta ključni tudi interpretabilnost in kvantifikacija negotovosti. Analize pomembnosti značilk zagotavljajo, da modeli temeljijo na ekološko smiselnih dejavnikih (npr. obremenitvah s hranili, temperaturnih anomalijah), razpršenost ansambelskih napovedi in napovedni intervali pa zagotavljajo mere negotovosti, ki podpirajo odločanje ob upoštevanju tveganj. Ponavljajoči se premiki domene (zaradi sezonskih ciklov, nadgradenj senzorjev ali novih dogodkov) zahtevajo periodično ponovno učenje, prenosno učenje in neprekinjeno vrednotenje, da se skozi čas ohrani zanesljivo delovanje [4, 5, 7].

Vrste algoritmov

V okoljski analitiki se pogosto uporablja več družin klasifikacijskih algoritmov:

- Logistična regresija: linearni verjetnostni model, ki zagotavlja interpretabilne koeficiente za binarne ali večrazredne izide.
- Stroji podpornih vektorjev (SVM): klasifikatorji, ki temeljijo na ločevalnem robu in z jedri ločujejo razrede v linearnih ali nelinearnih prostorih.
- Odločitvena drevesa: hierarhični modeli, ki podatke razdelijo v razrede z zaporednimi pragovi na podlagi značilk.
- Naključni gozdovi in gradientno ojačevanje: ansambli odločitvenih dreves, ki povečujejo napovedno moč in zmanjšujejo varianco.
- Konvolucijske nevronske mreže (CNN): modeli globokega učenja, ki iz slikovnih ali spektralnih podatkov samodejno izluščijo hierarhične značilke.

Mere vrednotenja

Vrednotenje klasifikacijskih modelov v okoljski analitiki zahteva več kot le splošno natančnost, saj so ekološke nevarnosti pogosto redke, podatkovni nabori pa močno neuravnoteženi.

Natančnost

- Delež pravilno klasificiranih vzorcev med vsemi vzorci.
- Uporabna je, kadar so razredi uravnoteženi.

- Omejitev: pri neuravnoteženih podatkih je lahko zavajajoča (npr. 95 % varnih in 5 % nevarnih primerov: trivialni klasifikator, ki vedno napove »varno«, že doseže 95-odstotno natančnost).

Preciznost

- Delež napovedanih pozitivnih primerov, ki so dejansko pozitivni.
- Visoka preciznost pomeni malo lažnih alarmov.
- Pomembna je, kadar so lažno pozitivni rezultati dragi (npr. nepotrebni sanacijski ukrepi).

Občutljivost (priklic)

- Delež dejansko pozitivnih primerov, ki so pravilno prepoznani.
- Visoka občutljivost pomeni malo spregledanih nevarnosti.
- Ključna je pri okoljskem spremljanju, kjer imajo lahko lažno negativni rezultati (npr. spregledana razlitja nafte ali nezaznano onesnaženje) hude posledice.

Mera F1

- Harmonična sredina preciznosti in občutljivosti.
- Uravnoteži kompromis med zaznavanjem nevarnosti in preprečevanjem lažnih alarmov.
- Pogosto se uporablja pri neuravnoteženih podatkovnih naborih kot skupni povzetek uspešnosti.

ROC–AUC (karakteristika delovanja sprejemnika)

- Meri kompromis med deležem resnično pozitivnih in deležem lažno pozitivnih rezultatov pri različnih pragovih.
- Omogoča celovit vpogled v uspešnost klasifikatorja.
- Omejitev: pri izraziti neuravnoteženosti razredov je lahko preveč optimistična.

PR–AUC (preciznost–občutljivost)

- Zajema razmerje med preciznostjo in občutljivostjo pri različnih pragovih.
- Ko so anomalije redke, je informativnejša od ROC–AUC.
- Posebej pomembna je pri zaznavanju ekoloških tveganj, kjer so nevarni dogodki redki, vendar imajo velik vpliv.

Okoljski problemi

Številne temeljne okoljske izzive je mogoče obravnavati kot klasifikacijske probleme, pri čemer različni algoritmi ponujajo komplementarne prednosti [4, 5].

Vodni sistemi

Naloge: zaznavanje škodljivega cvetenja alg, razvrščanje onesnaženosti, klasifikacija trofičnega stanja.

Primerni algoritmi:

- Logistična regresija: interpretabilna pri kemičnih/fizikalnih napovednih spremenljivkah [2];
- Omejitev: slabše deluje pri nelinearnih učinkih.
- Naključni gozdovi/gradientno ojačevanje: zajemajo kompleksne interakcije v mešanih hidrokemijskih podatkovnih naborih [4];

Omejitev: zmanjšana interpretabilnost.

Atmosferski sistemi

Naloge: razvrščanje kakovosti zraka, zaznavanje perjanic industrijskega onesnaženja [4].

Primerni algoritmi:

- SVM: učinkoviti pri visokorazsežnih senzorskih podatkih [4];
  Omejitev: velika računska zahtevnost pri obsežni uporabi.
- Gradientno ojačevanje: visoka napovedna uspešnost pri neuravnoteženih podatkih o kakovosti zraka [5];

  Omejitev: zahteva skrbno nastavitev.

Gozdarstvo in raba tal

Naloge: kartiranje požarne ogroženosti, klasifikacija stresa vegetacije, odkrivanje krčenja gozdov [4, 5].

Primerni algoritmi:

- Odločitvena drevesa: preprosta in razložljiva pri določanju vegetacijskih pragov;

  Omejitev: nagnjenost k prenaučenju.
- Naključni gozdovi: robustna uspešnost pri raznovrstnih napovednih spremenljivkah rabe tal [5];

  Omejitev: manj pregledna logika odločanja.

- Konvolucijske nevronske mreže (CNN): idealne za posnetke daljinskega zaznavanja vegetacije/pokrovnosti tal;
  Omejitev: potrebujejo veliko podatkov.

Zdravje ekosistema

Naloge: odkrivanje invazivnih vrst na podlagi eDNA, razvrščanje kategorij tveganja za bolezni prostoživečih živali [5, 7].

Primerni algoritmi:

- Stroji podpornih vektorjev (SVM): zelo primerni za visokorazsežne genomske podatke ali podatke eDNA [4];
  Omejitev: manjša razložljivost.
- Logistična regresija: razložljiv izhodiščni model za dejavnike tveganja bolezni [4];
  Omejitev: slabša uspešnost pri zapletenih interakcijah.

Na vseh teh področjih izbira algoritma uravnoteža napovedno točnost, razložljivost in računsko učinkovitost; ansambelski modeli in modeli globokega učenja so uspešni pri kompleksnih, visokorazsežnih podatkih [5], linearni modeli in modeli na podlagi dreves pa zagotavljajo preglednost in robustnost pri preprostejših nalogah [4].

Odkrivanje nevarnih/anomalnih objektov

Okoljske anomalije so pogosto redki, nepričakovani in prostorsko omejeni dogodki, ki pomenijo veliko ekološko tveganje. Primeri vključujejo nenadne skoke koncentracije trdnih delcev, nenadno škodljivo cvetenje alg, širjenje invazivnih vrst ali pojav mikroplastike in težkih kovin v vodnih in kopenskih sistemih. Ključni izziv je izrazito neravnovesje razredov: nevarni dogodki lahko predstavljajo manj kot 5 % opazovanj, zato jih je z običajnimi klasifikatorji težko odkriti. Poleg tega so anomalije pogosto značilne za posamezno lokacijo, zato modeli, naučeni na enem območju, brez skrbnega vrednotenja morda niso posplošljivi na drugo območje.

Seznam metod

Za odkrivanje anomalij v okoljskih podatkih se uporablja več pristopov:

- Statistično odkrivanje osamelcev: prepoznava skrajne vrednosti ob predpostavkah glede porazdelitve.

- Ponovno vzorčenje in stroškovno občutljivo učenje: ponovno uravnotežita porazdelitev razredov z nadvzorčenjem, podvzorčenjem ali določitvijo višjih kazni za lažno negativne rezultate.
- Ansambelsko učenje: naključni gozdovi, gradientno ojačevanje in zloženi modeli združujejo različne osnovne modele, da izboljšajo odpornost proti šumu.
- Odkrivanje anomalij z globokim učenjem: modeli na podlagi konvolucijskih mrež ali avtoenkoderjev zajamejo nelinearne vzorce in izpostavijo nenavadne prostorske ali spektralne značilnosti.

Primeri posebnih metod

Tri reprezentativne tehnike ponazarjajo, kako različni pristopi v praksi obravnavajo odkrivanje anomalij:

- Enorazredni SVM: nauči se odločitvene meje, ki obdaja »normalne« podatke, vzorce zunaj tega območja pa označi kot anomalije.
  Prednosti: učinkovit v visokorazsežnih prostorih, odporen proti zmernemu šumu.
  Omejitve: občutljiv na izbiro jedra in hiperparametre; računsko zahteven pri velikih podatkovnih naborih.
  Uporaba: odkrivanje neobičajnih epizod kakovosti zraka, prepoznavanje invazivnih vrst v vzorcih eDNA [6, 7].
- Naključni gozd: ansambel odločitvenih dreves, ki zajame zapletene interakcije med značilkami in na podlagi stopnje zaupanja v razvrstitev izpostavi osamelce.
  Prednosti: dobro razširljiv, odporen proti prenaučenju, zmožen obravnavati raznovrstne vhodne podatke.
  Omejitve: manj razložljiv kot preprosta drevesa; uspešnost je odvisna od nastavitev hiperparametrov in uravnoteženosti vzorca.
  Uporaba: odkrivanje onesnaženih tal na podlagi kemijskih analiz, prepoznavanje nenavadnih profilov kakovosti vode [3].
- Avtoenkoderji: nevronske mreže, naučene za rekonstrukcijo vhodnih podatkov; anomalije se odkrijejo, kadar je napaka rekonstrukcije velika.
  Prednosti: zajamejo kompleksne nelinearne odvisnosti, primerni za neprekinjeno spremljanje.
  Omejitve: za učenje potrebujejo dovolj »normalnih« podatkov, so računsko zahtevni in manj razložljivi.

Uporaba: odkrivanje anomalij v realnem času pri spremljanju kakovosti vode [1].

Študija primera: analiza tveganja na podlagi razširjenosti vrste z uporabo podatkov o pojavljanju iz GBIF in bioklimatskih napovednih spremenljivk WorldClim

Cilj

Ta študija primera ponazarja ponovljiv potek dela za odkrivanje okoljskega tveganja na podlagi modeliranja razširjenosti vrst (SDM). Cilj je prepoznati okoljske razmere, povezane s prisotnostjo ciljne vrste, in izdelati model strojnega učenja, ki oceni verjetnost pojavljanja na celotnem obravnavanem območju. V okviru ekološkega tveganja je mogoče območja z visoko napovedano primernostjo obravnavati kot prednostna območja za spremljanje in preprečevanje, zlasti pri invazivnih vrstah ali prenašalcih bolezni.

Potek dela je izveden v Pythonu samo s standardnimi knjižnicami za podatkovno znanost in odprto dostopnimi viri podatkov.

Uporabljene knjižnice

Uporabili bomo majhen nabor razširjenih Pythonovih knjižnic za pridobivanje podatkov, obdelavo podatkov, vzorčenje rastrskih podatkov, strojno učenje in vizualizacijo. Izbrani tehnološki nabor je namenoma preprost in temelji na »klasičnih« paketih, ki se pogosto uporabljajo v uporabni podatkovni znanosti. V nadaljevanju so navedene uporabljene knjižnice in njihov namen:

- **requests -** uporablja se za dostop do spletnih API-jev (GBIF) in prenos podnebnih slojev (WorldClim).
- **pandas** in **numpy -** uporabljata se za čiščenje in filtriranje podatkov ter inženirstvo značilk.
- **rasterio -** uporablja se za branje rastrskih datotek GeoTIFF in pridobivanje okoljskih spremenljivk za geografske koordinate.
- **scikit-learn** - zagotavlja klasifikacijske modele in metrike vrednotenja.
- **matplotlib -** uporablja se za preproste grafikone in interpretacijo modela.

Ta izbira podpira preglednost in omogoča preprosto izvajanje poteka dela v standardnih okoljih, kot sta Jupyter Notebook ali Google Colab.

Vir podatkov

V tej praktični vadnici se osredotočamo na vrsto ***Vespa velutina*** (azijski sršen), invazivno vrsto, ki se je razširila v več evropskih državah ter lahko ogroža lokalno biotsko raznovrstnost in čebelarstvo, saj pleni opraševalce, kot so medonosne čebele. Zato je ugotavljanje, kje se bo vrsta verjetno pojavila v prihodnje, praktičen primer odkrivanja okoljskega tveganja. Cilj vadnice je izdelati preprost model strojnega učenja, ki se nauči okoljskih razmer, povezanih s prisotnostjo vrste, nato pa za nove lokacije izračuna oceno primernosti habitata oziroma tveganja. Natančneje, podatkovni nabor sestavljajo zapisi o prisotnosti vrste *Vespa velutina* na izbranem območju in v izbranem časovnem obdobju (2000–2024), ki vključujejo koordinate zemljepisne širine/dolžine in metapodatke o opazovanju ter so bili preneseni prek javnega API-ja GBIF. Za pripravo vhodnih značilk je vsaka točka pojavljanja vrste dopolnjena s podnebnimi napovednimi spremenljivkami, pridobljenimi iz bioklimatskih spremenljivk WorldClim v2.1. To so globalni rastrski sloji, ki povzemajo dolgoročne vzorce temperature in padavin ter se pogosto uporabljajo pri modeliranju razširjenosti vrst. Končni podatkovni nabor tako vsako lokacijo poveže z nizom podnebnih kazalnikov (npr. povprečno letno temperaturo, sezonskostjo padavin) in omogoča nalogo binarnega razvrščanja na podlagi točk prisotnosti ter točk psevdoodsotnosti (ozadnih točk), vzorčenih na istem območju.

Viri podatkov: API GBIF Occurrence https://techdocs.gbif.org/en/openapi/v1/occurrence in bioklimatski podatki WorldClim v2.1 https://www.worldclim.org/data/worldclim21.html

Vrste podatkov

Vadnica združuje dve pogosti vrsti geoprostorskih podatkov, ki se pri okoljskih analizah pogosto uporabljata skupaj: *točkovne podatke o pojavljanju* (kje je bilo nekaj opaženo) in *rastrske podatke* (prostorske sloje, ki opisujejo okoljske razmere po celotnem območju).

- **Podatki o pojavljanju** (točkovna opazovanja): Prvi podatkovni nabor je tabela zapisov GBIF o pojavljanju vrste Vespa velutina. V kodi je vsako opazovanje shranjeno kot ena vrstica z geografskimi koordinatami (decimalLatitude, decimalLongitude) in osnovnimi metapodatki: gbifID, year, basisOfRecord, datasetKey ter (kadar je na voljo) coordinateUncertaintyInMeters. Te točke predstavljajo lokacije prisotnosti (oznaka 1), torej kraje, kjer je bila vrsta dejansko zabeležena. Ker GBIF večinoma zagotavlja le podatke o prisotnosti, ti zapisi ne vključujejo neposredno potrjenih odsotnosti.

```
# ------------------------------
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

- **Rastrski podatki** (okoljski sloji v mreži): Drugi podatkovni nabor sestavljajo bioklimatski rastri WorldClim v2.1, shranjeni kot datoteke GeoTIFF. Vsak raster je mreža, ki pokriva območje, vsaka celica pa vsebuje podnebno vrednost. V tej vadnici uporabljamo šest napovednih spremenljivk: BIO1, BIO4, BIO5, BIO12, BIO14, BIO15, ki bodo postale stolpci bio1, bio4, bio5, bio12, bio14 in bio15. Rastrske vrednosti se pridobijo z uporabo knjižnice 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 sta vrsti podatkov združeni (pridobivanje rastrskih vrednosti za točke): V koraku povezovanja vsako točko GBIF dopolnimo z rastrskimi vrednostmi: za vsako koordinato (lon, lat) preberemo vrednost iz vsakega rastra WorldClim in jo dodamo v nov stolpec, ki predstavlja značilko. Tako nastane standardna tabela za strojno učenje, v kateri je vsaka vrstica lokacija, vsaka napovedna spremenljivka pa podnebna spremenljivka.

```
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

Ker GBIF večinoma zagotavlja le zapise o prisotnosti, vadnica ustvari **točke ozadja** na istem območju in jim dodeli oznako 0. Te točke služijo kot približek lokacijam brez prisotnosti, kar omogoča podatkovni nabor za binarno razvrščanje.

```
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

Čiščenje podatkov

Podatkovni nabori o pojavljanju pogosto zahtevajo osnovno čiščenje, saj lahko vsebujejo podvojene koordinate (ista lokacija je navedena večkrat), negotove koordinate z majhno prostorsko natančnostjo in v redkih primerih točke zunaj izbranega območja zanimanja. Ker je cilj te vadnice povezati okoljske razmere s prisotnostjo vrste, bi lahko nenatančni ali ponovljeni zapisi pristranili analizo in nastali model strojnega učenja. Zato potek dela uporabi standardne filtre: odstrani zapise z manjkajočo zemljepisno širino ali dolžino, obdrži samo opazovanja znotraj omejitvenega pravokotnika območja zanimanja (ROI), odstrani podvojene pare koordinat in, kadar je podatek na voljo, po želji izloči točke z zelo veliko negotovostjo koordinat.

Končni podatkovni nabor za učenje modela

Točke prisotnosti in točke ozadja, za katere so bile pridobljene napovedne spremenljivke BIO, so združene v en podatkovni nabor za učenje. Matrika vhodnih značilk X vsebuje izbrane bioklimatske spremenljivke, pridobljene iz WorldClim (napovedne spremenljivke BIO), ciljni vektor y pa vsebuje binarne oznake, ki označujejo prisotnost oziroma psevdoodsotnost. Tako nastane standardni podatkovni nabor za nadzorovano učenje, primeren za učenje klasifikacijskih modelov.

```
# --------------------------------
# 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

Razdelitev na učno in testno množico ter izhodiščni model (logistična regresija)

Za oceno sposobnosti modela za posploševanje na nove podatke je podatkovni nabor razdeljen na učno in testno množico. Uporabljena je stratificirana delitev, ki v obeh podmnožicah ohrani delež točk prisotnosti in psevdoodsotnosti ter tako prepreči, da bi bile ocene uspešnosti pristranske zaradi neuravnoteženosti razredov. **Logistična regresija** je uporabljena kot prvi izhodiščni model, ker je preprosta in interpretabilna, zagotavlja verjetnostne napovedi ter je uveljavljena referenčna metoda za probleme binarne klasifikacije. **Mere vrednotenja:** Ker so ekološki podatkovni nabori o prisotnosti pogosto neuravnoteženi, je uspešnost modela ocenjena z merami, ki so informativnejše od same natančnosti, med njimi ROC–AUC, preciznost in občutljivost, mera F1 ter matrika zmede. To omogoča smiselnejšo razlago napovedne uspešnosti v okviru odkrivanja okoljskih tveganj.

```
# ------------------------------
# 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))
```

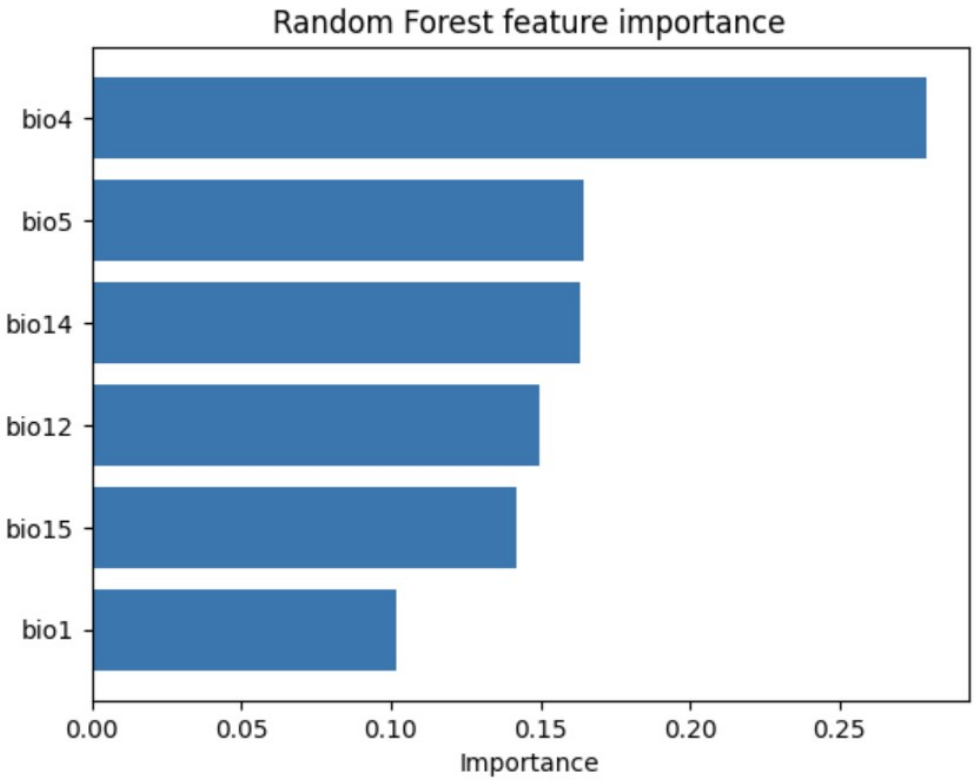
```
LogReg ROC-AUC: 0.9246512131422616
              precision    recall  f1-score   support

           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]]
```

Naključni gozd in pomembnost značilk

Naključni gozd uvedemo kot drugi model, ker so odnosi med vrsto in okoljem pogosto nelinearni ter lahko vključujejo interakcije med napovednimi spremenljivkami. Pri tabelaričnih okoljskih podatkih je navadno uspešen in dobro obravnava zapletene meje odločanja. Pomembnost značilk pokaže, katere podnebne spremenljivke najbolj prispevajo k ločevanju točk prisotnosti od psevdoodsotnosti. To podpira ekološko interpretacijo, na primer ugotavljanje povezav s toplejšimi razmerami, občutljivosti na padavine v sušnem obdobju ali omejitev zaradi sezonskosti. Korak zato podpira »razložljivo odkrivanje tveganja«.

```
# ------------------------------
# 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]]
```



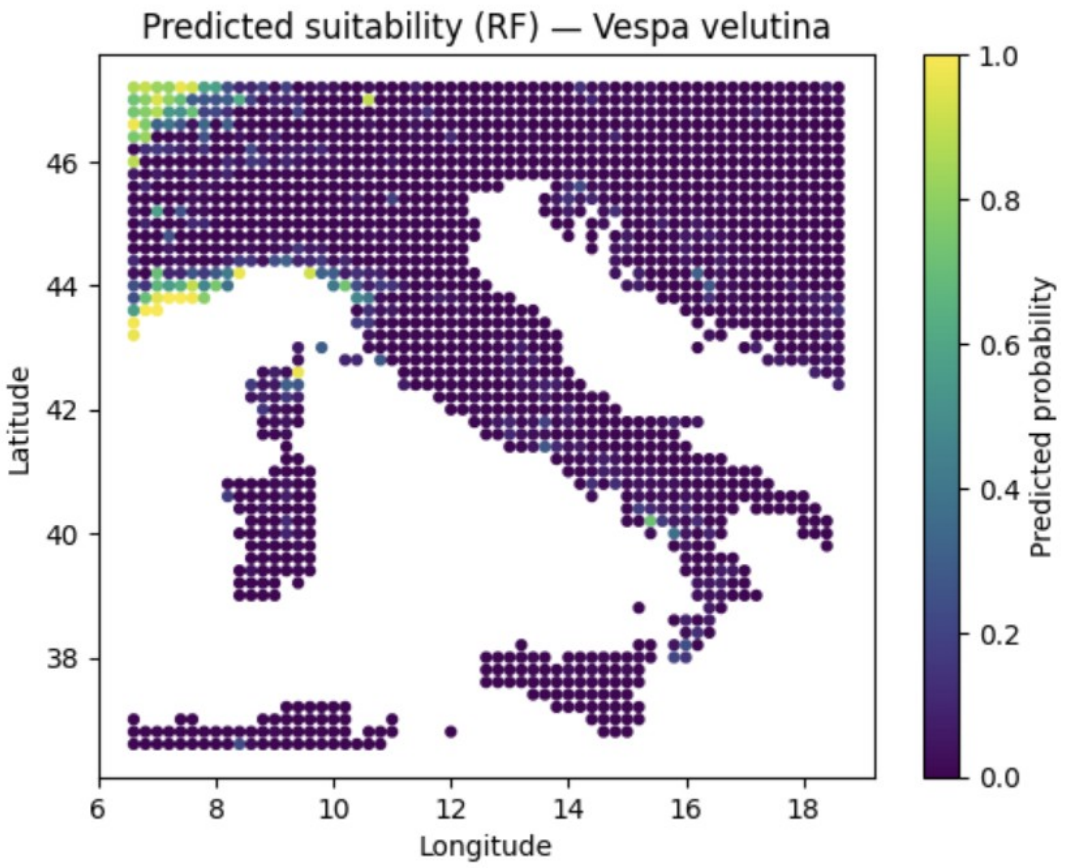
Model naključnega gozda na tem podatkovnem naboru dosega zelo visoko uspešnost, saj znaša ROC–AUC 0,99, splošna natančnost pa 0,96, kar kaže, da podnebne napovedne spremenljivke jasno ločujejo točke prisotnosti od točk psevdoodsotnosti. Matrika zmede kaže le omejeno število napak (26 lažno pozitivnih in 28 lažno negativnih primerov), poročilo o klasifikaciji pa potrjuje visoko preciznost in občutljivost za oba razreda. Graf pomembnosti značilk kaže, da je BIO4 (sezonskost temperature) najvplivnejša spremenljivka, kar pomeni, da je spremenljivost temperature skozi leto ključna za razlikovanje med primernimi in neprimernimi razmerami za vrsto Vespa velutina. Druge pomembne napovedne spremenljivke vključujejo BIO5 (najvišja

temperatura najtoplejšega meseca) in spremenljivke, povezane s padavinami (BIO12, BIO14, BIO15), medtem ko je BIO1 (povprečna letna temperatura) sama po sebi manj informativna. Rezultati podpirajo domnevo, da primernost okolja za vrsto *Vespa velutina* določajo ne le povprečne podnebne vrednosti, temveč tudi sezonskost in skrajne vrednosti, kar je skladno z ekološkimi pričakovanji glede ustalitve invazivnih vrst.

Prostorska napoved

Ključna prednost modeliranja razširjenosti vrst je, da se lahko naučeni model uporabi na lokacijah brez vzorcev. Za ponazoritev je na območju zanimanja ustvarjena mreža točk, ki jih ovrednoti prilagojeni klasifikator, s čimer nastane ploskev prostorske napovedi. Napovedano verjetnost je mogoče razlagati kot oceno primernosti habitata, približek verjetnosti ustalitve ali kazalnik prednostne obravnave pri spremljanju. Visoke napovedane vrednosti ustrezajo območjem, kjer bi se vrsta lahko ohranila, tudi če tam še ni bilo poročil o njenem pojavljanju.

```
# ----------------------------------
# 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()
```



Naloge za vajo

- Potek dela iz te vadnice ponovite za drugo vrsto po svoji izbiri (na primer za invazivno vrsto, opraševalca ali vrsto, pomembno za vašo regijo).
- V kodi zamenjajte SPECIES_NAME, prilagodite ROI svoji regiji, ohranite enako časovno obdobje (2000–2024) in izvedite celoten postopek: prenesite zapise o pojavljanju iz GBIF (točke prisotnosti), očistite koordinate, ustvarite točke psevdoodsotnosti, pridobite iste napovedne spremenljivke WorldClim (bio1, bio4, bio5, bio12, bio14, bio15) ter naučite logistično regresijo in naključni gozd. Pri izbiri primerne vrste preverite, ali poizvedba GBIF vrne dovolj zapisov (po čiščenju vsaj ~200 točk).
- V kratko poročilo vključite: (i) število uporabljenih točk prisotnosti, (ii) ROC–AUC in matriko zmede za oba modela ter (iii) graf pomembnosti značilk za naključni gozd.

- Nazadnje odgovorite na naslednja vprašanja: Kateri model je uspešnejši in zakaj? Katera spremenljivka BIO je najpomembnejša za vašo vrsto? Ali rezultati kažejo, da na vrsto bolj vplivajo omejitve, povezane s temperaturo, ali omejitve, povezane s padavinami?

Ocenjevanje

Kaj je glavni cilj uporabe klasifikacijskih modelov pri odkrivanju okoljskih tveganj?

A. Napovedovanje zveznih okoljskih spremenljivk, kot sta temperatura ali koncentracija onesnaževala
B. Združevanje okoljskih opazovanj v neoznačene gruče
C. Razvrščanje opazovanj v vnaprej določene kategorije tveganja (npr. nevarno oziroma varno)
D. Zmanjševanje razsežnosti okoljskih podatkovnih naborov

Zakaj splošna natančnost pogosto ne zadošča za vrednotenje klasifikacijskih modelov v okoljskih aplikacijah?

A. Ker okoljski podatki vedno vsebujejo šum
B. Ker natančnosti ni mogoče izračunati pri večrazrednih problemih
C. Ker so okoljski podatkovni nabori pogosto močno neuravnoteženi, nevarni dogodki pa so redki
D. Ker natančnost ne upošteva interpretabilnosti modela

Katera mera vrednotenja je najpomembnejša, kadar je cilj čim bolj zmanjšati število spregledanih okoljskih nevarnosti (lažno negativnih primerov)?

A. Preciznost
B. Natančnost
C. Občutljivost (priklic)
D. ROC–AUC

Kateri algoritem je posebej primeren za odkrivanje naftnih razlitij ali morskih odpadkov na satelitskih posnetkih ali posnetkih z droni?

A. Logistična regresija
B. Stroji podpornih vektorjev

C. Naključni gozdovi
D. Konvolucijske nevronske mreže (CNN)

Katera trditev najbolje opisuje vlogo izhodov klasifikacijskih modelov pri obvladovanju okoljskih tveganj?

A. V celoti nadomestijo strokovno presojo
B. Zagotavljajo povsem točne napovedi prihodnjih okoljskih razmer
C. Podpirajo odločanje s prepoznavanjem in prednostnim razvrščanjem kategorij tveganja
D. Odpravijo negotovost v sistemih spremljanja

Dodatni viri

Uvodne vadnice za napovedovanje s Pythonom (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.