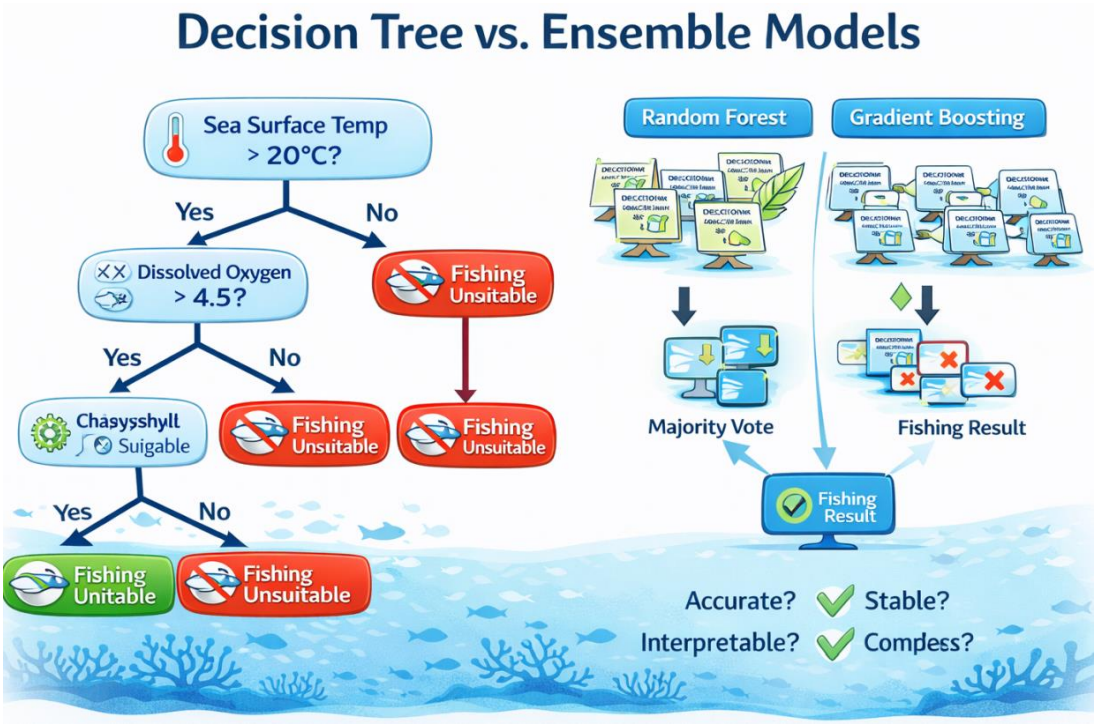
Druhý ročník 2025/2026	Druhé kolo	Úloha 3

Silnejšie modely: ako zlepšiť predikciu rybolovu?

V predchádzajúcej úlohe ste použili rozhodovací strom na predikciu vhodnosti rybolovu. Zistili ste, že strom dokáže vytvárať zrozumiteľné pravidlá, ale zároveň môže robiť chyby a byť citlivý na konkrétne dáta.



Predstavte si teraz, že by ste sa nespoliehali len na jeden model, ale na **viacero modelov naraz**. Každý z nich by sa učil trochu inak a robil by trochu iné chyby. Ak ich vhodne skombinujeme, môžeme získať presnejší a stabilnejší výsledok.

Práve na tomto princípe fungujú **ensemble metódy**. Ensemble modely (napr. Random Forest alebo Gradient Boosting) kombinujú viacero rozhodovacích stromov:

- každý strom sa učí na trochu iných dátach alebo iným spôsobom,
- jednotlivé stromy hlasujú alebo sa postupne opravujú,
- výsledný model je presnejší a menej citlivý na šum v dátach.

Napríklad Random Forest vytvorí desiatky až stovky stromov a ich rozhodnutia spriemeruje. Tým sa znižuje riziko, že model bude robiť chyby kvôli náhode alebo špecifickým dátam.

Vašou úlohou bude:

- porovnať jednoduchý rozhodovací strom s ensemble modelmi,
- zistiť, či sa zlepší presnosť,
- a zamyslieť sa nad tým, čo stratíme (interpretovateľnosť).

DATASET:

Použite rovnaký dataset ako v predchádzajúcej úlohe:
https://ics.upjs.sk/~antoni/marine_fishing_dataset.csv
Použite upravené dáta z Úlohy 2 vrátane nového atribútu fishing_suitable. Pri trénovaní modelov nepoužívajte stĺpec aquamaps_probability ako vstupný atribút, pretože z neho bol odvodený nový atribút fishing_suitable. Stĺpce sample_id a species taktiež nepoužívajte ako vstupy.

ÚLOHY:

a) Tréning ensemble modelov

Natrénujte aspoň dva z nasledujúcich modelov:

- Random Forest: n_estimators = 100, max_depth = 5
- Gradient Boosting: n_estimators = 100
- Bagging: n_estimators = 50

Použite rovnaké rozdelenie dát ako v Úlohe 2.

b) Vyhodnotenie modelov

Pre každý model určte:

- accuracy
- macro F1-score
- confusion matrix

Porovnajte tieto výsledky s rozhodovacím stromom z Úlohy 2.

c) Stabilita modelu

Model natrénujte aspoň trikrát s rôznym random_state a porovnajte variabilitu výsledkov (accuracy alebo F1-score).

d) Dôležitosť atribútov

Pre vybraný ensemble model:

- zobrazte feature importance (napríklad pre Random Forest)
- určte 3–5 najdôležitejších atribútov

Porovnajte ich s atribútmi zo stromu v Úlohe 2.

e) Interpretácia

Odpovedzte:

- ktorý model dosahuje najlepšiu presnosť?
- ktorý model je najlepšie vysvetliteľný?
- aký model by ste použili v praxi a prečo?

f) Diskusia

Zamyslite sa:

- prečo sú ensemble modely presnejšie než jeden strom?
- prečo je ich interpretácia náročnejšia?

Poznámky pre riešenie úloh druhého kola:

Pri riešení môžete používať internet. Môžete pracovať v ľubovoľnom softvéri: Excel, Google Sheets, Python, R alebo iba ručne na papieri. S prípadnými otázkami sa na nás môžete kedykoľvek obrátiť. Riešenia úlohy (dokumentácia + prípadný zdrojový kód) môžete odovzdať v **.zip priečinku** v termíne do **24.05.2026** cez formulár zverejnený na stránke https://vucap-challenge.science.upjs.sk/

Riešenia jednotlivých podúloh vhodne okomentujte, ak je to vhodné pridajte aj obrázky. Je možné odovzdať aj čiastočné riešenia jednotlivých úloh. Pri veľmi zaujímavom či prepracovanom riešení (pod)úlohy vám môžu byť udelené aj bonusové body.