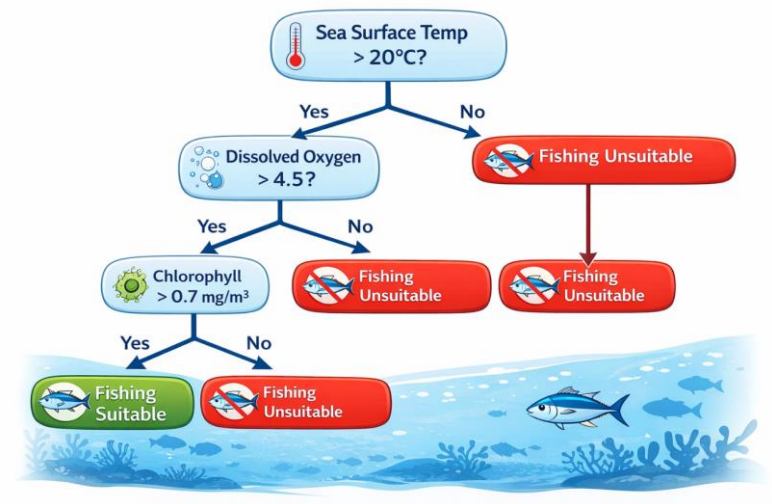
Druhý ročník 2025/2026	Druhé kolo	Úloha 2

Rozhodovanie pomocou stromu: kedy je vhodný rybolov?

Predstavte si, že ste súčasťou tímu, ktorý plánuje rybársku expedíciu v oceáne. Každý výjazd lode stojí tisíce eur: palivo, posádka, čas. Ak sa rozhodnete zle a vyplávate do oblasti, kde sa ryby nevyskytujú, znamená to veľkú stratu. Naopak, ak dokážete správne odhadnúť, kde sú podmienky vhodné, môžete výrazne zvýšiť úspešnosť rybolovu.



Problém je, že oceán je veľmi komplexné prostredie. **Výskyt rýb** (napríklad tuniaka) závisí od viacerých faktorov:

- teploty vody,
- množstva kyslíka,
- dostupnosti potravy (planktónu),
- prúdenia oceánu,
- hĺbky a vzdialenosti od pobrežia.

Tieto faktory spolu vytvárajú zložité vzťahy, ktoré človek len ťažko odhadne „od oka“. Práve tu prichádza na rad **strojové učenie**.

Rozhodovací strom je model strojového učenia, ktorý sa rozhoduje podobne ako človek pomocou postupnosti otázok. Na začiatku stojí koreň stromu a model sa postupne „pýta“ otázky na jednotlivé atribúty, napríklad:

- Je teplota vody väčšia ako 20 °C?
- Je množstvo chlorofylu väčšie ako 1?

Podľa odpovedí sa presúva po vetvách stromu až k listu, ktorý obsahuje finálne rozhodnutie (napr. vhodné / nevhodné na rybolov). Každá cesta od koreňa po list predstavuje jedno pravidlo, napríklad:

ak teplota > 20 a chlorofyl > 1, tak je to vhodné na rybolov

Výhodou rozhodovacích stromov je, že:

- sú ľahko pochopiteľné
- dokážu vysvetliť svoje rozhodnutia
- pracujú aj s nelineárnymi vzťahmi medzi premennými

V tejto úlohe použijete rozhodovací strom, ktorý:

- automaticky nájde pravidlá v dátach,
- dokáže odpovedať na otázky typu: „Za akých podmienok je rybolov vhodný?“
- a hlavne: svoje rozhodnutia vie vysvetliť.

Takéto pravidlá sú veľmi cenné, pretože ich vedia pochopiť aj odborníci mimo informatiky (napr. biológovia alebo rybári). Vašou úlohou bude vytvoriť model, ktorý:

- predpovie vhodnosť rybolovu,
- nájde zrozumiteľné pravidlá,
- a vysvetlí svoje chyby.

DATASET:

Použite dataset zo stránky:
https://ics.upjs.sk/~antoni/marine_fishing_dataset.csv
Dataset obsahuje približne:

- 1000 riadkov (vzorky z rôznych častí oceánu)
- geografické súradnice (latitude, longitude)
- environmentálne atribúty oceánu
- cieľovú premennú vyjadrujúcu pravdepodobnosť výskytu tuniaka

Dáta vychádzajú z modelu výskytu tuniaka (AquaMaps),

Dataset obsahuje stĺpec:
aquamaps_probability – pravdepodobnosť výskytu tuniaka (hodnota od 0 do 1).
Vašou úlohou bude túto pravdepodobnosť previesť na dve hodnoty (0 alebo 1), tak že

vytvorte nový atribút **fishing_suitable**:
fishing_suitable = 1, ak pravdepodobnosť ≥ 0.6 (vhodné na rybolov)
fishing_suitable = 0, ak pravdepodobnosť < 0.6 (nevhodné)
Môžete experimentovať aj s iným prahom (napr. 0.7 alebo 0.8) a sledovať, ako sa menia výsledky modelu.

Identifikačné

- sample_id – identifikátor záznamu
- species – druh (Thunnus obesus – tuniak)

Geografické atribúty:

- lat, lon – geografická poloha
- ocean_basin – oceán (Tichý, Atlantický, Indický)

Fyzikálne a biologické atribúty:

- sea_surface_temp_c – teplota povrchovej vody
- salinity_psu – slanosť
- chlorophyll_mg_m3 – množstvo chlorofylu (indikátor planktónu)
- primary_production_mgC_m3_day – primárna produkcia
- dissolved_oxygen_mmol_m3 – množstvo rozpusteného kyslíka
- current_velocity_m_s – rýchlosť prúdenia
- mixed_layer_depth_m – hĺbka miešanej vrstvy
- thermocline_depth_m – hĺbka termoklíny
- distance_to_land_km – vzdialenosť od pevniny
- sea_ice_fraction – pokrytie ľadom
- occupancy_depth_m – typická hĺbka výskytu tuniaka

ÚLOHY:

a) Príprava dát

- Načítajte dataset.
- Najprv vytvorte cieľovú premennú fishing_suitable zo stĺpca aquamaps_probability.
- Overte kvalitu dát (chýbajúce hodnoty, extrémy).
- Rozdeľte dáta na tréningovú a testovaciu časť (napr. 80:20).
- Použite stratifikované rozdelenie.

b) Tréning modelu

Natrénujte model rozhodovacích stromov (Decision Tree):

- model 1: **plytký strom** (napr. s maximálnou hĺbkou max_depth = 3)
- model 2: **hlbší strom** (napr. max_depth = 6 alebo max_depth = None)

Pri trénovaní modelu stĺpec aquamaps_probability už nepoužívajte medzi vstupnými atribútmi, pretože z neho bola cieľová premenná odvodená. Stĺpce sample_id a species taktiež nepoužívajte ako vstupy do modelu.

c) Vyhodnotenie

Pre oba modely určte:

- accuracy
- macro F1-score
- confusion matrix

d) Vizualizácia

- vykreslite strom
- identifikujte najdôležitejšie atribúty

e) Interpretácia

Nájdite aspoň **5 pravidiel**, napr.:
ak oxygen > 5 a chlorophyll > 1, tak je to vhodné na rybolov

f) Analýza chýb

Vyberte 3 chybné predikcie z testovacej množiny a pri každej uveďte:

- skutočnú triedu,
- predikovanú triedu,
- rozhodovaciu cestu v strome alebo kľúčové atribúty.

g) Porovnanie

- rozdiel medzi plytkým a hlbokým stromom.

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.