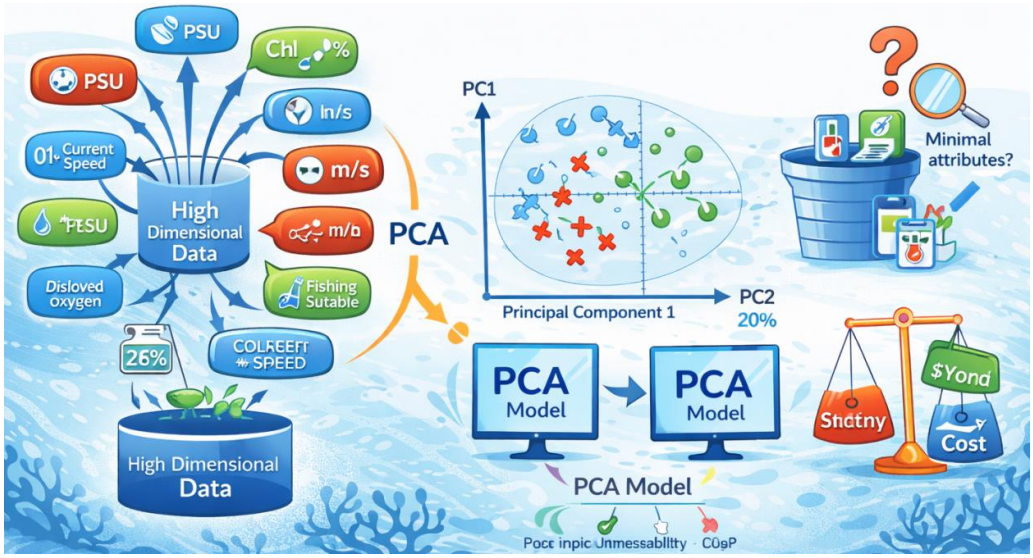
Druhý ročník 2025/2026	Druhé kolo	Úloha 4

Skryté vzory v dátach: redukcia dimenzie

V predchádzajúcich úlohách ste pracovali s množstvom atribútov, ktoré opisujú oceánske prostredie. Každý z nich nesie určitú informáciu, no nie všetky sú rovnako dôležité. Niektoré môžu byť navzájom podobné alebo redundantné. Predstavte si, že máte k dispozícii desiatky senzorov, ktoré merajú rôzne veličiny v oceáne. Každý senzor stojí peniaze, vyžaduje údržbu a môže sa pokaziť. Preto je dôležité vedieť:

- ktoré informácie sú naozaj kľúčové
- a či vieme počet atribútov znížiť bez veľkej straty presnosti.



Jedným z nástrojov, ktorý nám s tým pomáha, je **PCA (Principal Component Analysis)**. PCA dokáže:

- zredukovať počet premenných,
- zachytiť najdôležitejšie vzory v dátach,
- a premietnuť dáta do menšieho počtu nových „komponentov".

Tieto komponenty už nie sú pôvodné fyzikálne veličiny (napr. teplota alebo chlorofyl), ale ich kombinácie.

Vašou úlohou bude:

- preskúmať štruktúru dát pomocou PCA,
- zistiť, či sa triedy dajú oddeliť v 2D priestore,
- porovnať modely na pôvodných a redukovaných dátach,
- a navrhnúť čo najjednoduchší model s minimom atribútov.

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.

ÚLOHY:

a) Preskúmanie dát

- preskúmajte vzťahy medzi atribútmi (napr. korelácie)
- identifikujte podobné alebo redundantné premenné

b) PCA analýza

- Pred aplikáciou PCA najprv štandardizujte dáta, napr. pomocou StandardScaler a tiež odstráňte stĺpce aquamaps_probability, sample_id a species a použite iba numerické atribúty.
- Aplikujte PCA na tréningové dáta a následne použite rovnakú transformáciu na testovacie dáta.
- Pre vizualizáciu použite 2 komponenty, pre model môžete použiť viac (napr. podľa vysvetlenej variability).

c) Interpretácia PCA

Okomentujte:

- či sú triedy oddelené alebo sa prekrývajú (uvedomte si, že PCA neberie do úvahy triedy, preto oddelenie nemusí byť výrazné)
- identifikujte oblasti, kde sa triedy prekrývajú.
- koľko variability vysvetľujú prvé 2 komponenty

d) Model na PCA dátach

- natrénujte model (Decision Tree alebo Random Forest) na PCA komponentoch
- vyhodnoťte accuracy, macro F1-score

Porovnajte s modelmi z predchádzajúcich úloh (porovnajte, či redukcia dimenzie znižuje alebo zlepšuje výkon modelu).

e) Minimalistický model

Predstavte si, že máte len obmedzený počet senzorov.

- vyberte 3–5 atribútov (napr. najdôležitejšie podľa modelu)
- natrénujte nový model

Vyhodnoťte:

- ako sa zmenila accuracy
- ktoré atribúty sú najdôležitejšie
- či je jednoduchší model stále použiteľný

f) Diskusia

Odpovedzte:

- pomáha PCA zlepšiť model?
- aká je výhoda a nevýhoda PCA z pohľadu interpretácie?
- oplatí sa použiť menej atribútov v praxi?

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.