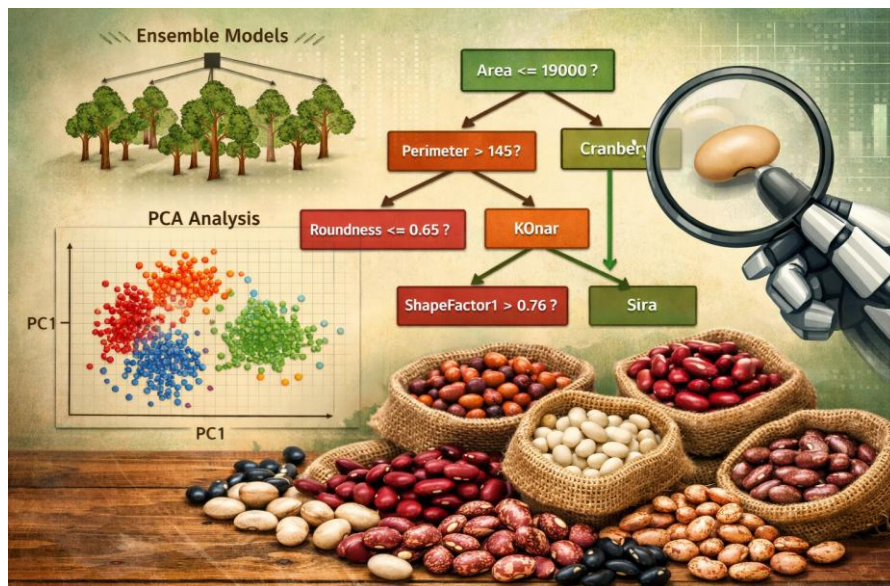


Úloha 1: Rozpoznávanie odrôd fazule pomocou vysvetliteľných stromových modelov

V poľnohospodárstve, potravinárstve aj pri automatizovanom triedení plodín je dôležité vedieť rýchlo a spoľahlivo rozlíšiť rôzne druhy semien. V praxi sa na to často používajú obrazové alebo geometrické charakteristiky objektov získané z kamier a senzorov. Pri takýchto úlohách však nestačí iba dosiahnuť vysokú presnosť. Dôležité je aj to, aby model vedel svoje rozhodnutie **vysvetliť**: ktoré vlastnosti boli rozhodujúce, aké pravidlá viedli ku klasifikácii a kde sa model mýli.



V tejto úlohe budete pracovať s reálnym datasetom **Dry Bean** z UCI Machine Learning Repository. Dataset obsahuje 13 611 vzoriek siedmich odrôd suchej fazule. Pre každé zrnko je k dispozícii 16 numerických príznakov, odvodených z obrazovej analýzy tvaru a veľkosti semien. Dataset vznikol na základe snímok z vysoko rozlišovacej kamery a je voľne dostupný na stiahnutie: <https://archive.ics.uci.edu/dataset/602/dry%2Bbean%2Bdataset>

Vašou úlohou bude vytvoriť a analyzovať systém, ktorý:

- klasifikuje odrodu fazule pomocou rozhodovacích stromov,
- porovná jeden strom so skupinami stromov (ensemble trees),
- využije PCA na vizualizáciu a analýzu dát,
- a čo najlepšie vysvetlí, prečo model rozhoduje práve takto.

Príprava dát

- Stiahnite si dataset a načítajte ho do svojho programu.
- Overte, či sa v dátach nachádzajú chýbajúce alebo podozrivé hodnoty.
- Rozdeľte dáta na tréningovú a testovaciu časť, napríklad v pomere 80:20.
- Použite **stratifikované rozdelenie**, aby zostalo zastúpenie tried podobné v oboch častiach.
- Ak je to vhodné, pripravte aj štandardizovanú verziu dát pre PCA.

Preskúmanie dát a PCA

- Zistite, ktoré príznaky (atribúty) spolu silno súvisia.
- Použite metódu **PCA** na zníženie dimenzie dát aspoň na 2 hlavné komponenty.
- Vytvorte 2D graf vzoriek po projekcii do PCA priestoru.
- Okomentujte:
 - ktoré triedy sú dobre oddelené,
 - ktoré triedy sa prekrývajú,
 - koľko variability vysvetľujú prvé dve alebo tri hlavné komponenty.

Tréning rozhodovacieho stromu

- Natrénujte klasifikačný **Decision Tree** model.
- Vyskúšajte aspoň dve nastavenia:
 - jednoduchý, plytký strom,
 - hlbší alebo menej obmedzený strom.
- Porovnajajte napríklad vplyv parametrov:
 - max_depth,
 - min_samples_leaf,
 - criterion (gini vs. entropy).
- Vyhodnoťte model pomocou metrík:
 - accuracy,
 - macro F1-score,
 - confusion matrix.
- Vizualizujte finálny strom alebo jeho časť.

Interpretácia stromu

Z rozhodovacieho stromu vyťažte čo najviac informácií. Nájdite aspoň **5 zaujímavých pravidiel**, napríklad v štýle:

Ak je Area > ... a ShapeFactor1 < ..., potom ide pravdepodobne o odrodu ...

Vyberte aspoň **3 nesprávne klasifikované vzorky** a vysvetlite:

- akou cestou strom prešiel,
- v ktorom bode rozhodovania sa model „pomýlil“,
- ktoré triedy si boli podobné.

Porovnajajte:

- jednoduchší strom s nižšou presnosťou, ale lepšou interpretovateľnosťou,
- zložitejší strom s vyššou presnosťou, ale horšou čitateľnosťou.

Skupiny stromov (ensemble)

Natrénujte aspoň **dva ensemble modely stromov**, napríklad:

- Random Forest,
- Gradient Boosting,
- alebo Bagging s rozhodovacími stromami.

Porovnajajte ich s jedným rozhodovacím stromom z hľadiska:

- presnosti,
- macro F1-score,
- stability výsledkov,
- interpretovateľnosti.

Okomentujte, prečo skupiny stromov často dosahujú lepšie výsledky než jeden strom, ale zároveň sa horšie vysvetľujú.

PCA vs. stromové modely

Porovnajajte:

- modely tréňované na **pôvodných príznakoch (originálnych atribútoch)**,
- modely tréňované na **PCA komponentoch (po aplikácii PCA)**.

Vyhodnoťte:

- či PCA pomohla alebo zhoršila výsledky,
 - či sa po PCA zlepšila vizualizácia, ale zhoršila interpretovateľnosť,
 - či stromové modely vôbec potrebujú PCA.
-

Vysvetliteľný model zrníek fazule

Predstavte si, že firma chce vytvoriť jednoduché zariadenie, ktoré vie merať len niekoľko vlastností zrnka, nie všetkých 16.

Navrhните postup:

- vyberte len **3 až 5 najdôležitejších príznakov**,
- natrénujte na nich nový strom,
- porovnajte jeho výsledok s modelom využívajúcim všetky príznaky.

Na záver odpovedzte:

- Aké minimum meraní ešte stačí na rozumné triedenie?
- O koľko sa zhorší presnosť?
- Oplatí sa jednoduchší a lacnejší model?

Čo odovzdať

1. Kód riešenia vo forme .py alebo .ipynb.
2. Stručnú dokumentáciu alebo komentáre v kóde.
3. Výstupy modelov a grafy:
 - PCA vizualizáciu,
 - confusion matrix,
 - porovnanie metrík modelov,
 - vizualizáciu aspoň jedného rozhodovacieho stromu.
4. Krátky textový komentár k výsledkom, v ktorom vysvetlíte:
 - čo bolo najdôležitejšie pre klasifikáciu,
 - ktoré triedy sa plietli,
 - aký bol rozdiel medzi jedným stromom a skupinami stromov,
 - čo ukázalo PCA.

Hodnotenie

0 – 100 bodov

Pri hodnotení sa prihliada na:

- **30 bodov** – správnosť prípravy dát a implementácie,
 - **20 bodov** – kvalitu modelovania pomocou rozhodovacích stromov,
 - **15 bodov** – porovnanie so skupinovými modelmi stromov,
 - **15 bodov** – zmysluplné využitie PCA,
 - **20 bodov** – interpretáciu výsledkov, komentár, nápaditosť a zrozumiteľnosť.
-