Tema projektnog zadatka:
Binarna klasifikacija jestivih i nejestivih gljiva uz odabir optimalnog klasifikatora i interpretaciju značajki

Ideja i cilj zadatka
Cilj projekta je analizirati skup podataka o gljivama dostupnih na poveznici Kaggle Mushroom Classification, gdje svaka gljiva nosi oznaku "jestiva" ili "nejestiva". Primjenom različitih algoritama binarne klasifikacije (K-neighbors, Random Forest, Decision Tree, Logistic Regression itd.) vrednovat će se njihova učinkovitost prema relevantnim metrikama (F1, ROC-AUC, točnost i dr.) s ciljem identifikacije najuspješnijeg klasifikatora za ovu domenu.
Dodatno, bit će provedena analiza značajki (feature selection) kako bi se utvrdilo koje karakteristike najviše utječu na klasifikaciju jestivosti gljiva. Na temelju odabranih značajki ponovit će se trening klasifikacijskih modela radi dodatne procjene promjena u njihovoj efikasnosti.

Planirana metodologija

- Učitavanje i predobrada podataka (one-hot encoding, rukovanje nedostajućim vrijednostima, balansiranje klasa ako je potrebno).
- Trening i vrednovanje klasifikatora (KNN, Random Forest, Decision Tree, Logistic Regression) koristeći k-fold cross-validation.
- Mjerenje performansi uz F1, ROC-AUC, accuracy i confusion matrix.
- Primjena metoda selekcije značajki (npr. Random Forest Feature Importance, Recursive Feature Elimination).
- Ponovno treniranje klasifikatora na reduciranom skupu značajki te usporedba rezultata s početnim performansama.

Potencijalni izazovi i ograničenja

- Skup podataka je kategoriziran, pa obrada i kodiranje varijabli može biti izazovna (posebno kod algoritama osjetljivih na dimenzionalnost).
- Moguća pojava slabo reprezentiranih klasa ili značajki koje ne doprinose klasifikaciji.
- Rizik od prenaučenosti kod kompleksnijih modela (poput Random Forest), osobito kod korištenja svih značajki.
- Interpretacija dobivenih rezultata može biti ograničena ako model koristi velik broj značajki s malim individualnim doprinosom.
- Potencijalna nesigurnost u generalizaciji modela na druge, realne podatke izvan danog skupa.