Predikce akademického stresu – oponentura

Volba modelovacích algoritmů

Autoři zvolili dva klasifikační modely – Logistic Regression a Random Forest Classifier. Volba je zcela adekvátní vzhledem k typu úlohy (binární klasifikace), malé velikosti datasetu (140 řádků) i charakteru proměnných.

Hyperparameter tuning

Ladění hyperparametrů pomocí GridSearchCV je implementováno bez problémů:

- **logreg**: ladí se C, class_weight, penalizace je pevně nastavena na L2
- **random forest**: ladí se n_estimators, max_depth, min_samples_split, min_samples_leaf

Hyperparametry jsou vypsány, nejlepší modely jsou použity (např. logreg má C=0.1, class_weight=balanced).

Replikovatelnost výsledků

- Kód obsahuje kompletní pipeline od načtení dat až po evaluaci.
- Preprocessing je řešen pomocí ColumnTransformer a Pipeline, tedy reprodukovatelně.
- Je správně uveden explicitní random_state u train_test_split

Volba evaluačních metrik a interpretace výsledků

Autoři používají:

- accuracy
- ROC AUC
- confusion matrix
- classification report (precise/recall/f1)

To je vhodná sada pro binární klasifikaci a zvlášť pro nevyvážená data – dataset má 89 : 51. Výběr AUC je proto na místě.

Autoři správně upozorňují, že logistická regrese dosahuje nízké schopnosti rozlišovat mezi třídami.

Souhrn

Projekt je kvalitně zpracovaný, s dobře navrženým postupem a vhodnou volbou modelů (Logistic Regression a Random Forest). Preprocessing je proveden korektně pomocí pipeline a ColumnTransformeru, což zajišťuje čistotu a reprodukovatelnost. Hyperparameter tuning i evaluace pomocí accuracy, AUC a confusion matrix jsou provedeny správně. Pdf report je přehledný, jasný, srozumitelný a bez jazykových chyb.

Chybí analýza feature importance. U Random Forestu by bylo vhodné ukázat, které proměnné jsou nejvýznamnější nebo zda některé nejsou redundantní

Chybí závěr typu „Doporučení do praxe“:

- co lze predikcí získat
- jak by se dal model prakticky využít
- co je největší přínos

Vysvětlení kroků je místy příliš stručné

- Proč je použita medianová imputace?
- Proč jen Logistic Regression a Random Forest?
- Proč se stres binarizuje zrovna 1–3 vs. 4–5?
- Proč je accuracy hlavní metrika?