Instrukcje do sprawozdania

Student ma za zadanie wyznaczyć główne składowe z analizy PCA, a następnie zbudować model regresji liniowej na otrzymanych głównych składowych. Wykorzystane zostaną dane z ostatniego ćwiczenia (dane-leki.xlsx). Student powinien:

- Zaimportować dane, oraz podzielić je na zmienne zależne i niezależne
- Wyznaczyć główne składowe dla X (*fit_transform*)
- Podzielić dane na zbiór treningowy i testowy przy pomocy *train_test_split(test_size=0.33, random_state=42)*
- Wybrać optymalną ilość składowych na podstawie wykresu zależności RMSE od ilości głównych składowych z uwzględnieniem walidacji krzyżowej K-Fold*(n_splits=10, shuffle=True, random_state=1)*
- Zbudować model z optymalną ilością głównych składowych i wyznaczyć wartości statystyczne modelu: R^2,Q^2,RMSE,$RMSE_{ex}$
- [Na dodatkowe 0,5pkt na następnej wejściówce] Na podstawie wykresu ładunków czynnikowych określić wkład poszczególnych zmiennych w główne składowe.
- Krótko zinterpretować otrzymane wyniki

Wnioski i interpretację można uwzględnić jako komentarze w kodzie lub wysłać oddzielnie plik z kodem i wnioski w pdf. Sprawozdanie należy wysłać na adres: rafal.ziniewicz@phdstud.ug.edu.pl

Użyteczne biblioteki : sklearn–metrics, sklearn–linear regression, sklearn.model_selection - KFold, cross_val_score, train_test_split, sklearn.decomposition-PCA, numpy, pandas, matplotlib, seaborn, statsmodels.api