Projekt 1

Projekty można realizować w parach lub samodzielnie. Każda para musi mieć unikalny w obrębie roku zestaw (zestaw danych, algorytm, optymalizowany parametr). Proszę o przesłanie list z wybranymi tematami najpóźniej do 14 marca, w przeciwnym razie 15 marca przygotuję losowy przydział tematów.

Projekty należy oddawać na platformie UPEL do 28 marca (w przeciwnym razie z projektu wystawiona będzie ocena 2.0).

1. Jeden z zestawów danych:
    a. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer (klasyfikacja, brakujące dane).
    b. https://archive.ics.uci.edu/ml/datasets/Car+Evaluation (klasyfikacja).
    c. https://archive.ics.uci.edu/ml/datasets/Echocardiogram (klasyfikacja, brakujące dane).
    d. https://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival (klasyfikacja, brakujące dane).
    e. https://archive.ics.uci.edu/ml/datasets/Census+Income (klasyfikacja, brakujące dane, dość duży zbiór).
    f. https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations (klasyfikacja)
    g. https://archive.ics.uci.edu/ml/datasets/Automobile (regresja).
    h. https://archive.ics.uci.edu/ml/datasets/Bias+correction+of+numerical+prediction+model+temperature+forecast (regresja, brakujace dane)
    Uwaga: zbiór trzeba przerobić. Next_Tmax i Next_Tmin to są błędne przewidywania istniejącego systemu (można je wykorzystać jako cechy), natomiast poprawne wartości należy brać z kolumn Present_Tmax i Present_Tmin dla kolejnego dnia (i tej samej stacji).
    i. https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset (regresja). Można ograniczyć się do wersji zagregowanej po dniach.
    j. https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise (regresja).
    k. https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime (regresja).
    l. Inne z repozytorium UCI: https://archive.ics.uci.edu/ml/datasets.php (należy unikać zbiorów z cechami typu tekstowego czy szeregów czasowych, zbyt dużych zbiorów (>50 000 próbek), zbyt małych zbiorów (<100 próbek) oraz zbiorów gdzie liczba próbek jest mniejsza niż liczba cech).
2. Algorytm uczenia maszynowego:
    a. Klasyfikacja:
        i. SVM,
        ii. Lasy drzew decyzyjnych (RandomForestClassifier).
        iii. C4.5

        iv. Regresja logistyczna.
    b. Regresja:
        i. Drzewa regresji (DecisionTreeRegressor).
        ii. ElasticNet.
        iii. Regresja wielomianowa
3. Sposób walidacji: 10-krotna walidacja krzyżowa
4. Optymalizowany parametr:
    a. klasyfikacja:
        i. Accuracy (dokładność).
        ii. Macierz pomyłek (należy różnym błędnym klasyfikacjom przypisać różne wagi).
        iii. Sensitivity (czułość) -- dla klasyfikacji binarnej.
        iv. Precision -- dla klasyfikacji binarnej.
        v. AUC -- dla klasyfikacji binarnej.
    b. Regresja:
        i. Błąd średniokwadratowy.
        ii. Średni błąd bezwzględny.
        iii. Ułamek wyjaśnianej wariancji (explained_variance_score).

Brakujące dane: wystarczy SimpleImputer

W raporcie należy zamieścić:
1. Krótki opis zestawu danych: liczba cech i ich typy, czy występują brakujące dane, rodzaj problemu (klasyfikacja, regresja), liczba instancji (próbek).
2. Krótki opis wybranej metody uczenia maszynowego (około 2-3 zdania) + opis parametrów.
3. Sposób wyboru zbioru testowego.
4. (Na >= 4.0): opis działania metody wyboru hiperparametrów

Szacowana długość raportu: od 1 do 3 stron A4.

Ocena:
- Na 3.0: działający model uczenia maszynowego, przetestowano i porównano kilka hiperparametrów; policzenie wybranego optymalizowanego parametru. Dla klasyfikacji narysowanie macierzy pomyłek a dla regresji krzywej uczenia: https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html#sphx-glr-auto-examples-model-selection-plot-learning-curve-py
- Na 4.0: zastosowano prawidłowo walidację krzyżową do znalezienia optymalnych hiperparametrów (wybrać 2) na siatce (grid search). Policzone wybranego optymalizowanego parametru na zbiorze testowym dla optymalnego klasyfikatora i narysowanie dla niego macierzy pomyłek lub krzywej uczenia.
- Na 5.0: Zbadanie wpływu normalizacji, standaryzacji i PCA (na cechach będących liczbami rzeczywistymi) na proces uczenia (dodanie ich jako trzeci optymalizowany

hiperparametr o pięciu wartościach: brak normalizacji czy standaryzacji, normalizacja, standaryzacja, dwa warianty PCA z różnymi wyborami liczby głównych składowych).