ZADANIE PROJEKTOWE

I. Treść zadania

Wybierz jeden z listy tematów. W każdym temacie zgadują się trzy zbiory danych. Dwa z nich dotyczą problemu klasyfikacji, jeden z kolei problemu regresji. Twoim zadaniem jest analiza każdego zbioru za pomocą minimum trzech różnych algorytmów uczenia maszynowego oraz wykonanie badania ich działania w zależności od ich parametrów wejściowych. Dla każdego zbioru wskaż najlepszy spośród wybranych algorytmów oraz wartości najlepszych parametrów.

II. Efekt pracy

Efektem pracy ma być sprawozdanie (wydruk) składające się z następujących części:

1. Dane studenta (imię i nazwisko, numer indeksu, temat projektu, data oddania pracy).
2. Opis zbiorów danych.
3. Użyte algorytmy wraz z krótkim opisem teoretycznym i opisem działania.
4. Przygotowanie danych (czyszczenie danych, normalizacja, standaryzacja, kodowanie zmiennych kategorycznych, podział na zbiory treningowe i testowe).
5. Wyniki klasyfikacji i regresji (wykresy, tabele, porównania metryk).
6. Analiza wpływu parametrów (jakie parametry zostały przetestowane, jak zmienia się wynik, jaki parametr był optymalny).
7. Wizualizacje wyników (wykresy skuteczności od wartości testowanych parametrów).
8. Podsumowanie i wnioski.

III. Sugerowane algorytmy do wykorzystania

Dla klasyfikacji: LogisticRegression, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, SVC.

Dla regresji: LinearRegression, KNeighborsRegressor, DecisionTreeRegressor, Ridge, SVR.

Opisy algorytmów można znaleźć na stronie: https://scikit-learn.org/stable/

IV. Lista tematów wraz z linkami do zbiorów danych

Temat 1: Klasyfikacja zdrowia i ceny domów

- Klasyfikacja 1: Breast Cancer Wisconsin (Diagnostic) Dataset – UCI
  https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

- Klasyfikacja 2: Digits Dataset – scikit-learn
  https://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html

- Regresja: California Housing Dataset – scikit-learn
  https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset

Temat 2: Klasyfikacja marketingowa i zarobki

- Klasyfikacja 1: Titanic Dataset – Kaggle
  https://www.kaggle.com/c/titanic

- Klasyfikacja 2: Social Network Ads Dataset – Kaggle
  https://www.kaggle.com/datasets/rakeshrau/social-network-ads

- Regresja: Salary Dataset – Kaggle
  https://www.kaggle.com/datasets/abhishek14398/salary-dataset-simple-linear-regression?utm_source=chatgpt.com

Temat 3: Klasyfikacja genetyczna i analiza medyczna

- Klasyfikacja 1: Wine Dataset – UCI
  https://archive.ics.uci.edu/dataset/109/wine

- Klasyfikacja 2: Pima Indians Diabetes Dataset – Kaggle
  https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

- Regresja: Diabetes Dataset – scikit-learn
  https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset

Temat 4: Klasyfikacja obrazów i regresja nieruchomości

- Klasyfikacja 1: Fashion MNIST Dataset – GitHub
  https://github.com/zalandoresearch/fashion-mnist

- Klasyfikacja 2: Make Moons Dataset – scikit-learn
  https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html

- Regresja: Ames Housing Dataset – Kaggle
  https://www.kaggle.com/datasets/prevek18/ames-housing-dataset

Temat 5: Klasyfikacja tekstu i predykcja filmów

- Klasyfikacja 1: 20 Newsgroups Dataset – scikit-learn
  https://scikit-learn.org/stable/datasets/real_world.html#newsgroups-dataset

- Klasyfikacja 2: SMS Spam Collection Dataset – Kaggle
  https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

- Regresja: MovieLens Small Dataset – GroupLens
  https://grouplens.org/datasets/movielens/

Temat 6: Klasyfikacja klientów i zużycie energii

- Klasyfikacja 1: Bank Marketing Dataset – UCI
  https://archive.ics.uci.edu/dataset/222/bank+marketing

- Klasyfikacja 2: Heart Disease Dataset – UCI
  https://archive.ics.uci.edu/dataset/45/heart+disease

- Regresja: Energy Efficiency Dataset – UCI
  https://archive.ics.uci.edu/dataset/242/energy+efficiency

Temat 7: Klasyfikacja finansowa i przewidywanie kredytu

- Klasyfikacja 1: German Credit Data – UCI
  https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

- Klasyfikacja 2: Banknote Authentication Dataset –
  https://archive.ics.uci.edu/dataset/267/banknote+authentication

- Regresja: Student Performance Dataset – UCI
  https://archive.ics.uci.edu/dataset/320/student+performance

Temat 8: Klasyfikacja zachowań użytkowników

- Klasyfikacja 1: Adult Income Dataset – UCI
  https://archive.ics.uci.edu/dataset/2/adult

- Klasyfikacja 2: Online Shoppers Intention Dataset – UCI
  https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

- Regresja: Bike Sharing Dataset – UCI
  https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset