

**Warsaw University
of Technology**



Wstęp do Uczenia Maszynowego



Zasady punktacji

Aby zaliczyć laboratorium:

- **min. 31/60** pkt z obu projektów
- **min. 11/30** pkt z każdego z projektów
- Projekty są realizowane w grupach 2-osobowych.
- Kody i raport należy wysłać mailowo do 23:59 w dniu oddania (jan.sawicki@pw.edu.pl)
- Każdy rozpoczęty tydzień po terminie to -50% punktów za dany termin
- Prezentację wygłasza min. 1 osoba z zespołu
- Brak prezentacji lub kodu skutkuje brakiem punktów za cały kamień milowy.
- **Kamień milowy 1** (5 pkt)
 - Prezentacja (max. 10 min)
 - Kod (notebook)
- **Kamień milowy 2** (5 pkt)
 - Prezentacja (max. 10 min)
 - Kod (notebook)
- **Kamień milowy 3** (20 pkt)
 - Prezentacja (max. 10 min)
 - własny projekt (10 z 20 pkt)
 - walidacja projektu innej grupy (10 z 20 pkt)
 - Raport (opisany notebook)

Projekt - klasyfikacja

1. Kamień milowy 1 - Eksploracyjna Analiza Danych

- a. Wyznaczenie celu biznesowego – podział zmienne objaśniające i zmienną objaśnianą
- b. Analiza statystyczna i wizualizacja – obliczenie podstawowych miar statystycznych (np. średnia, mediana, odchylenie standardowe), wykresy (np. histogramy, wykresy pudełkowe, macierze korelacji).
- c. Identyfikacja braków danych i anomalii – wykrywanie brakujących wartości, analiza wartości odstających i ich potencjalny wpływ na model.
- d. Badanie zależności między zmiennymi – analiza korelacji między cechami, wykrywanie redundancji, eksploracja rozkładów zmiennych względem etykiety docelowej.

2. Kamień milowy 2 - Inżynieria Cech

- a. Selekcja cech – wybór najbardziej informatywnych cech poprzez testy statystyczne, algorytmy selekcji, synergie międzycechowe
- b. Tworzenie nowych cech – generowanie dodatkowych zmiennych na podstawie istniejących (np. interakcje między zmiennymi, transformacje matematyczne itp.
- c. Transformacja zmiennych kategoriycznych – one-hot encoding, target encoding, ordinal encoding w zależności od kontekstu.

3. Kamień milowy 3 - Modelowanie i Walidacja

- a. Dobór modelu – testowanie różnych algorytmów (np. regresja liniowa, drzewa decyzyjne) oraz ich dostrajanie
- b. Dostrajanie hiperparametrów – dostosowanie parametrów modelu za pomocą technik takich jak Grid Search czy Random Search.
- c. Podział zbioru danych i metody walidacji – wykorzystanie metod takich jak walidacja krzyżowa, utrzymanie zbioru testowego
- d. Ocena jakości modelu – wybór odpowiednich metryk w zależności od problemu (np. dokładność, F1-score, RMSE itp.)
- e. Walidacja projektu innej grupy na podstawie danych walidacyjnych (ok. 20% zbioru danych, na którym nie był trenowany model)
 - i. Druga grupa zapewnia dane i kod do weryfikacji

Projekt - klasteryzacja

1. Kamień milowy 1 - Eksploracyjna Analiza Danych

- a. Wyznaczenie celu biznesowego – podział zmienne objaśniające i zmienną objaśnianą, proporcja klas
- b. Dla obrazów: np. histogramy kolorów/rozmiarów/proporcji, analiza kanałów RGB, natężenie kolorów, wykrywanie anomalii, weryfikacja poprawności/uszkodzeń plików, inne
- c. Dla tekstów: np. długość dokumentów, słów, rozkład częstości słów (word frequency, TF-IDF), najczęściej występujące słowa i n-gramy, chmury słów, stemming, lematyzacja, złożoność (text complexity), inne

2. Kamień milowy 2 - Inżynieria Cech

- a. Redukcja wymiarów cech (PCA, t-SNE)
- b. Dla obrazów: np. zmiana rozmiaru i standaryzacja obrazów, skala szarości, augmentacja danych (odbicie lustrzane, rotacja, przycięcie, jasność), ekstrakcja cech (Haralick, Gabor), entropia, zanurzenia (CLIP, ViT, ResNet), inne
- c. Dla tekstu: np. tokenizacja, usuwanie stopwords, stemming, lematyzacja, TF-IDF, zanurzenia (Word2Vec, GloVe, BERT), inne

3. Kamień milowy 3 - Modelowanie i Walidacja

- a. Dobór algorytmu klasteryzacji – testowanie różnych metod (np. K-Means, DBSCAN, hierarchiczna klasteryzacja).
- b. Dostrajanie hiperparametrów – dostosowanie parametrów modelu, np. liczby klastrów (K-Means), min. liczba punktów (DBSCAN) itp.
- c. Ocena jakości klasteryzacji – dostosowanie parametrów algorytmu klasteryzacji za pomocą technik takich jak metoda łokcia, analiza silhouette, aby osiągnąć najlepszą separację i koherencję klastrów.
- d. Walidacja projektu innej grupy na podstawie danych walidacyjnych (ok. 20% zbioru danych, na którym nie był trenowany model)
 - i. Druga grupa zapewnia dane i kod do weryfikacji

| | Klasyfikacja | | Klasteryzacja | |
|--------------|--------------|------------------------|---------------|------------------------|
| Jestem grupą | Mam projekt | Waliduję projekt grupy | Mam projekt | Waliduję projekt grupy |
| 1 | 1 | 2 | 1 | 3 |
| 2 | 2 | 3 | 2 | 4 |
| 3 | 3 | 4 | 3 | 5 |
| 4 | 4 | 5 | 4 | 6 |
| 5 | 5 | 6 | 5 | 1 |
| 6 | 6 | 1 | 6 | 2 |

| | Klasyfikacja | | Klasteryzacja | |
|--------------|--------------|------------------------|---------------|------------------------|
| Jestem grupą | Mam projekt | Waliduję projekt grupy | Mam projekt | Waliduję projekt grupy |
| 1 | 1 | 2 | 1 | 3 |
| 2 | 2 | 3 | 2 | 4 |
| 3 | 3 | 4 | 3 | 5 |
| 4 | 4 | 5 | 4 | 6 |
| 5 | 5 | 6 | 5 | 7 |
| 6 | 6 | 7 | 6 | 1 |
| 7 | 7 | 1 | 7 | 2 |

| | Klasyfikacja | | Klasteryzacja | |
|--------------|--------------|------------------------|---------------|------------------------|
| Jestem grupą | Mam projekt | Waliduję projekt grupy | Mam projekt | Waliduję projekt grupy |
| 1 | 1 | 2 | 1 | 3 |
| 2 | 2 | 3 | 2 | 4 |
| 3 | 3 | 4 | 3 | 5 |
| 4 | 4 | 5 | 4 | 6 |
| 5 | 5 | 6 | 5 | 7 |
| 6 | 6 | 7 | 6 | 8 |
| 7 | 7 | 8 | 7 | 1 |
| 8 | 8 | 1 | 8 | 2 |

Projekt - klasyfikacja - zbiory danych

1. <https://www.kaggle.com/datasets/samayashar/fraud-detection-transactions-dataset>
2. <https://www.kaggle.com/datasets/adilshamim8/student-performance-and-learning-style>
3. <https://www.kaggle.com/datasets/aizahzeeshan/lung-cancer-risk-in-25-countries>
4. <https://www.kaggle.com/datasets/anthonytherrien/depression-dataset>
5. <https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube>
6. <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>
7. <https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>
8. <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package>

Projekt - klasteryzacja - zbiory danych

1. <https://www.kaggle.com/datasets/muratkokludataset/pistachio-image-dataset>
2. <https://www.kaggle.com/datasets/phucthaiv02/butterfly-image-classification>
3. <https://www.kaggle.com/datasets/anshtanwar/jellyfish-types>
4. <https://www.kaggle.com/datasets/aksha05/flower-image-dataset>
5. <https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification>
6. <https://www.kaggle.com/datasets/saurabhshahane/ecommerce-text-classification>
7. <https://www.kaggle.com/datasets/starblasters8/human-vs-llm-text-corpus>
8. <https://www.kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text>

DataCamp

- Opcjonalny
- Do uzyskania jest maks. **5 pkt** za **2 zaliczone kursy**
- https://www.datacamp.com/groups/shared_links/4a0456134c780c6afede0c2e9e93b8238da0d9b5132db9b4aeb6e9f3522d3897
⊘ Proszę nie udostępniać linku dalej ze względu na ograniczenie liczebności “grupy” DataCampowej