# UMA - projekt wstępny

Filip Ryniewicz, Miłosz Cieśla

3 stycznia 2025

*Techniki oceny klasyfikacji dla zestawów danych dotyczących raka piersi*

## 1 Opis projektu

Projekt będzie polegał na porównaniu technik ocen klasyfikacji na danych dotyczących raka piersi. Planujemy wytrenować 3 modele (które będą dokładnie opisane w kolejnym punkcie) i porównać ich osiągi za pomocą wybranych technik oceny klasyfikacji.

## 2 Ogólny opis wybranych algorytmów

Mamy zamiar wytrenować:

- Wielowarstwowy perceptron z funkcją aktywacji ReLU w warstwach ukrytych oraz funkcją sigmoid w warstwie wyjściowej. Zdecydowaliśmy się na MLP, ponieważ uważamy, że powinien on dobrze poradzić sobie z klasyfikacją binarną na wybranym dataset.
- Algorytm K-Najbliższych sąsiadów. Wybraliśmy go ze względu na jego prostotę i intuicyjność w wykrywaniu podobieństw między próbkami.
- Las losowy. Wybraliśmy go, ponieważ sprawdza się w klasyfikacji złożonych danych dzięki odporności na szum i dobrej generalizacji.

Użyjemy modeli z biblioteki PyTorch. Dodatkowo, dokonamy optymalizacji hiperparametrów w celu znalezienia wartości, dla których będziemy otrzymywać najlepsze wyniki dla każdego z modeli, pozwoli nam to na dokładne porównanie ich jakości za pomocą wybranych przez nas miar.

## 3 Szacunkowy plan eksperymentów

W ramach eksperymentów przeanalizujemy otrzymane z modeli wyniki, implementując następujące techniki ocen klasyfikacji:

1. Error rate - stosunek źle sklasyfikowanych próbek względem wszystkich. Pokazuje jak często model popełnia błędy.
2. Accuracy - stosunek prawidłowo sklasyfikowanych próbek względem wszystkich. Miara ta mówi, jak często model jest poprawny. Może być myląca na zbiorze danych o niezrównoważonej liczności klas, lecz uważamy, że warto ją sprawdzić, chociażby do porównania z innymi miarami.
3. True Positive Rate/Sensitivity/Recall - stosunek poprawnie sklasyfikowanych próbek pozytywnych do wszystkich, które powinny być sklasyfikowane jako pozytywne. Obrazuje, jak skutecznie model identyfikuje pozytywne pozytywne przypadki (rak złośliwy).

4. False positive rate - stosunek nieprawidłowo sklasyfikowanych próbek pozytywnych do wszystkich, które powinny być sklasyfikowane jako negatywne. Pokazuje, jak często model błędnie klasyfikuje negatywne próbki jako pozytywne, generując zarazem fałszywe alarmy.
5. Precision - stosunek sklasyfikowanych próbek pozytywnych do wszystkich sklasyfikowanych jako pozytywne. Mierzy dokładność modelu w klasyfikowaniu próbek jako pozytywne.
6. F1-Score - średnia harmoniczna precision i recall. Jest powrzechnie stosowany w problemach medycznych, gdzie próbki sklasyfikowane niepoprawnie ciągną za sobą poważne konsekwencje.
7. Analizę krzywą ROC - analiza zależności między True Positive Rate, a False Positive Rate. Pokazuje jak model radzi sobie z różnymi progami klasyfikacji.
8. AUC-ROC - pole pod krzywą ROC - mierzy jak dobrze model odróżnia klasy. Jest szczególnie użyteczny w problemach z niezrównoważonym zbiorem danych.

Dodatkowo, będziemy mierzyć czasy trenowania i predykcji wszystkich modeli.

# 4 Zbiór danych

Zbiór danych Breast Cancer Wisconsin (Diagnostic) zawiera 569 instancji zawierających 30 atrybutów uczących, ID oraz klasę. Próbki danych dzielą się na 2 klasy:

- **M** - Malignant breast cancer
- **B** - Benign breast cancer

Dane zostały wygenerowane na podstawie zdigitalizowanego obrazu biopsji cienkoigłowej (FNA) masy piersiowej pacjentek.

Atrybuty zawierają informacje o komórkach nowotworowych takie jak:

- promień
- tekstura
- obwód
- powierzchnia
- gładkość
- zwięzłość
- wklęsłość
- punkty wklęsłe
- symetria
- wymiar fraktalny