# Inteligencja Obliczeniowa - Projekt 1

Marek Ochocki, Łukasz Gosek

Kwiecień 2020

## 1 Zestaw danych

Wybrany przez nas zbiór danych jest jednym z popularniejszych w literaturze dotyczącej uczenia maszynowego. Został dostarczony przez Instytut Onkologii i zawiera informacje o pacjentkach ze zdiagnozowanym rakiem piersi.

W zbiorze znajduje się 286 instancji dwóch klas – 201 dla klasy 'no-recurrence-events' oraz 85 dla klasy 'recurrence-events'. Każda instancja opisana jest przez 9 cech różnych typów:

1. liczbowe – 'deg-malig'
2. liczbowe przedziałowe – 'age', 'tumor-size', 'inv-nodes' (traktowane przez nas jak zmienne kategoryczne)
3. logiczne – 'node-caps', 'irradiat'
4. kategoryczne – 'menopause', 'breast', 'breast-quad'

W zborze występują brakujące dane (dla cech 'node-caps' oraz 'breast-quad').

Rozpatrywany zbiór danych posłuży nam do rozwiązania problemu klasyfikacji.

## 2 Preprocessing

W celu ich eliminacji brakujących danych został wykorzystany SimpleImputer z biblioteki scikit-learn, wykorzystując strategię najczęściej występującej wartości (z racji obecności zmiennych kategorycznych).

Zmienne typu boolowskiego zostały zmapowane na liczby rzeczywiste – fałsz $\rightarrow 0$, prawda $\rightarrow 1$.

Zmienne kategoryczne oraz liczbowe przedziałowe zostały zakodowane przy użyciu One Hot Encoding (również z wykorzystaniem biblioteki scikit-learn).

Klasy zostały również zmapowane na wartości liczbowe - 'no-recurrence-events' → 0, 'recurrence-events' → 1.

W ten sposób uzyskano zestaw cech wyłącznie typu rzeczywistego.

# 3 Wybrany algorytm

Wybrany przez nas algorytm to Losowy Las Decyzyjny (Random Forest Classifier). Jest to zespołowa metoda uczenia maszynowego rozszerzająca koncepcję drzew decyzyjnych. Polega ona na tworzeniu wielu (często kilkuset lub nawet tysięcy) drzew, z których każdemu przydzielany jest podzbiór cech (lub wszystkie cechy – zależnie od zastosowanych parametrów) na podstawie których dokonują one predykcji. Do uczenia poszczególnych drzew często wykorzystuje się technikę bootstrap. Finalną predykcją lasu jest najczęściej występująca wartość predykcji poszczególnych drzew (dla klasyfikacji), lub ich średnia (dla regresji).

Przy korzystaniu z lasów losowych użytkownik może kontrolować wiele parametrów, lecz do najbardziej znaczących należą (nazewnictwo wg. implementacji w scikit-learn):

1. n_estimators – liczba drzew w lesie
2. max_depth – maksymalna głębokość dla każdego drzewa (niepodanie wartości skutkuje brakiem ograniczenia)
3. class_weight – wagi przypisywane poszczególnym klasom (szczególnie przydatny w przypadku niezbalansowanych zbiorów danych)
4. min_samples_split – minimalna liczba próbek wymagana do powstania rozgałęzienia w drzewie
5. min_samples_leaf – minimalna liczba próbek dla których tworzony jest liść drzewa

# 4 Wybór zbioru testowego

Zbiory treningowy i testowy zostały utworzone przy użyciu funkcji z biblioteki scikit-learn. Wykonano stratyfikowany podział na pomieszanych danych, rozmiar zbioru testowego to 20% wszystkich danych. Dzięki temu uzyskano

zbiory treningowy i testowy o takim samym rozkładzie instancji poszczególnych klas.

## 5 Metoda wyboru hiperparametrów

W toku przeprowadzonych eksperymentów zbadana została odpowiedź modelu na dane poddane różnym rodzajom preprocessingu (brak, normalizacja, standaryzacja oraz PCA z dwoma różnymi liczbami głównych składowych). Dla tych danych szukane były optymalne wartości parametrów modelu przy użyciu GridSearchCV z biblioteki scikit-learn.

Metoda ta przyjmuje słownik, w którym znajdują się dopuszczalne wartości dla poszczególnych parametrów modelu. Następnie, dla każdej możliwej kombinacji wartości parametrów, dokonuje uczenia modelu i jego ewaluacji przy użyciu walidacji krzyżowej (w naszym przypadku 10-krotnej). Optymalizowanym przez nas parametrem była średnia precyzja klasyfikacji. Zestaw parametrów dla którego ta metryka była najlepsza został uznany za optymalny.

Optymalizacji poddaliśmy trzy parametry: n_estimators, max_depth oraz class_weight (z racji dosyć niezbalansowanych klas). Najlepsza uzyskana precyzja wyniosła 71,2%, dla danych poddanych normalizacji, a optymalne wartości parametrów to: 'class_weight': None, 'max_depth': 3, 'n_estimators': 19.