Uczenie się Maszyn
Projekt "Zastosowanie drzew decyzyjnych do klasyfikacji miejsc rozcięcia w sekwencji DNA"

Dokumentacja końcowa

Igor Markiewicz

1 Spis treści

2 Założenia

- W trakcie procesu testowania programu okazało się że istnieje niewielka liczba przykładów w obu dostępnych zbiorach, które zawierają dodatkową literę N, oznaczającą najprawdopodobniej nieznaną lub dowolną zasadę azotową. Dlatego też do istniejących już czterech wartości atrybutów A, C, G i T została dołożona wartość N.
- Podział na zbiory uczący, walidacyjny i testowy przebiega w dwóch etapach. Najpierw procentowo określa się ile przykładów z pierwotnie dostępnego zbioru zostaje przeznaczonych na uczenie i walidację (reszta idzie na testowanie). Następnie z części przeznaczonej na uczenie i walidację, określa się procentowo ile przykładów jest przeznaczonych na uczenie (reszta idzie na walidację).
- Dla każdego pomiaru losowanie zbiorów, uczenie, (przycinanie) i testowania zostają przeprowadzone dziesięciokrotnie w celu poprawienia statystyki, a wyniki są wyznaczane średnią arytmetyczną (i ewentualnie zaokrąglane).
- Do przetestowania budowania drzewa został użyty przykład przedstawiony w dokumentacji wstępnej.
- Badania :
  - Eksperyment 1 - Podział zbioru pierwotnego na zbiór uczący o liczności 10 %, 20 %, ... , 90 % oraz zbiór testowy i badanie parametrów matrix confusion, błędu.
  - Eksperyment 2 - Podział zbioru pierwotnego na zbiór będący sumą zbioru uczącego i walidujacego (70 %) oraz na zbiór testowy (30 %), a następnie podział na zbiór uczący o liczności 10 %, 20 %, ... , 90 % i zbiór walidujący. Badamy parametry matrix confusion oraz błąd zarówno dla zbioru walidującego (czyli przed przycięciem drzewa) jak i dla zbioru testowego (po przycięciu drzewa, gdzie zbiorem do przycinania jest zbiór walidujący).

3 Przycinanie drzewa

3.1 Opis

Jako algorytm przycinania drzewa zostało wybrane Reduce Error Pruning, będące algorytmem zstępującym.

funkcja pruneDecisionTree(*root*, *data*)
argumenty wejściowe:

- *root* - korzeń drzewa decyzyjnego
- *data* - zbiór przycinający

zwraca: - (działa w miejscu względem oryginalnego drzewa);
1: **dopóki** korzeń nie jest sprawdzony
2: pruneEngine(*root*, *data*);
3: **koniec dopóki**

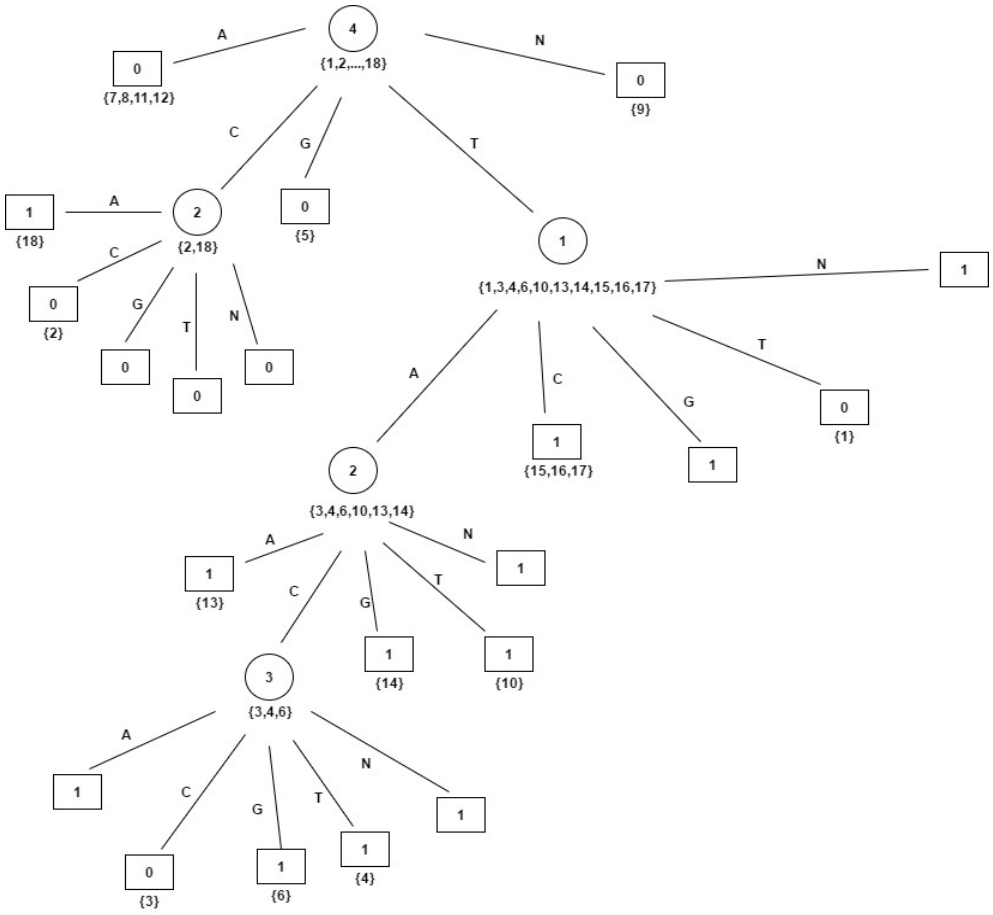
funkcja pruneEngine(*node*, *data*)
argumenty wejściowe:

- *node* - węzeł drzewa decyzyjnego
- *data* - zbiór przycinający

zwraca: - (działa w miejscu względem oryginalnego drzewa);
1: **jeśli** wszyscy potomkowie węzła nie są liśćmi lub nie są sprawdzeni lub nie ma superpozycji tych stanów **to**
2: **dla każdego** potomka węzła
3: **jeśli** potomek nie jest liściem i nie jest sprawdzony **to**
4: pruneEngine(*child*, *data*);
5: **koniec jeśli**
6: **koniec dla każdego**
7: **w przeciwnym przypadku**
8: ustaw węzeł jako sprawdzony;
9: sprawdź klasyfikację c_1 w podanym węźle drzewa i podanym zbiorze;
10: zbuduj przewidywaną klasyfikację c_2 na podstawie kategorii mniejszościowej w klasyfikacji c_1;
11: oblicz błąd dla c_1 i c_2;
12: **jeśli** błąd dla klasyfikacji c_2 jest mniejszy lub równy błędowi dla klasyfikacji c_1 **to**
13: ustaw węzeł jako liść;
14: przypisz do utworzonego liścia kategorię mniejszościową klasyfikacji c_1;
15: Usuń rekurencyjnie potomków i ustaw wskaźniki na dzieci na wartość domyślną;
16: **koniec jeśli**
17: **koniec w przeciwnym przypadku**
18: **koniec jeśli**

3.2 Przykład

id	x	d
1	TGAT	0
2	ACGC	0
3	ACCT	0
4	ACTT	1
5	ATAG	0
6	ACGT	1
7	TTTA	0
8	CTTA	0
9	AANN	0
10	ATTT	1
11	CACA	0
12	GTAA	0
13	AACT	1
14	AGCT	1
15	CTTT	1
16	CAAT	1
17	CGGT	1
18	AAAC	1

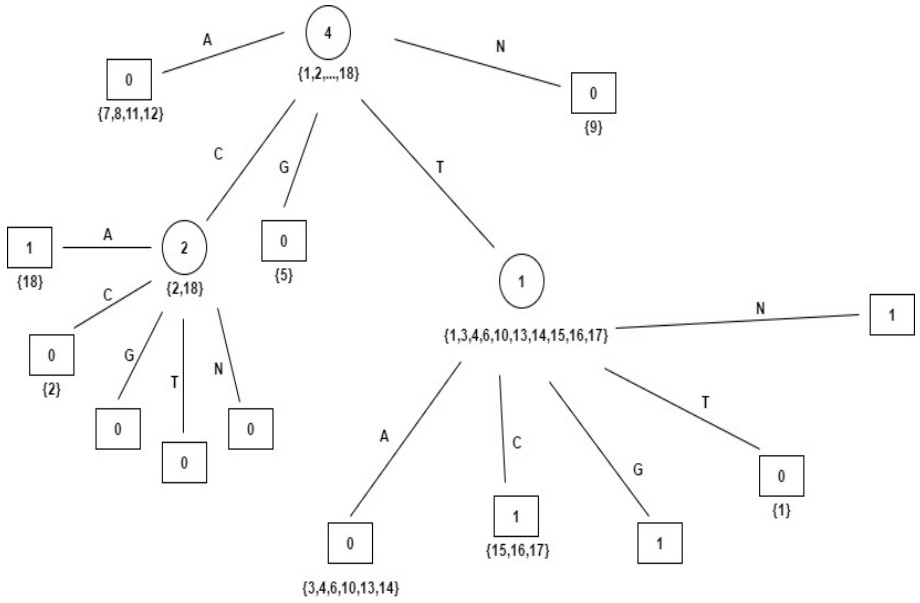


Rysunek 1: Zbudowane przez algorytm drzewo decyzyjne (bez randomizacji)

Przypuśćmy że chcemy sprawdzić jak algorytm zachowa się dla węzła o teście 2. (na ścieżce 4-1-2). W tym celu zamieniamy w zbiorze przykładów klasyfikację dla :

id	**x**	**d**
10	ATTT	0
13	AACT	0
14	AGCT	0

Błąd dla węzła : $\delta_{\text{node}} = \frac{3}{6} = \frac{1}{2}$
Błąd dla hipotetycznego liścia (o kategorii mniejszościowej 0): $\delta_{\text{leaf}} = \frac{2}{6} = \frac{1}{3}$
W efekcie algorytm przycinania dla danego węzła tworzy drzewo zgodne z naszymi oczekiwaniami :



Rysunek 2: Drzewo po przycięciu dla danego węzła

4 Analiza i wnioski

4.1 Schemat Tabeli

True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

4.2 Eksperyment 1.

4.2.1 Rozcięcia typu donorowego

- 10 %

TP = 700	FN = 303
FP = 309	TN = 3417

Błąd $\approx 12.95\%$

- 20 %

TP = 685	FN = 214
FP = 210	TN = 3094

Błąd $\approx 10.10\%$

- 30 %

TP = 611	FN = 172
FP = 159	TN = 2736

Błąd $\approx 9.03\%$

- 40 %

TP = 445	FN = 117
FP = 111	TN = 2337

Błąd $\approx 8.86\%$

- 50 %

TP = 361	FN = 87
FP = 89	TN = 1953

Błąd $\approx 8.71\%$

- 60 %

TP = 266	FN = 87
FP = 89	TN = 1564

Błąd $\approx 8.42\%$

- 70 %

TP = 266	FN = 65
FP = 67	TN = 1564

Błąd $\approx 8.41\%$

- 80 %

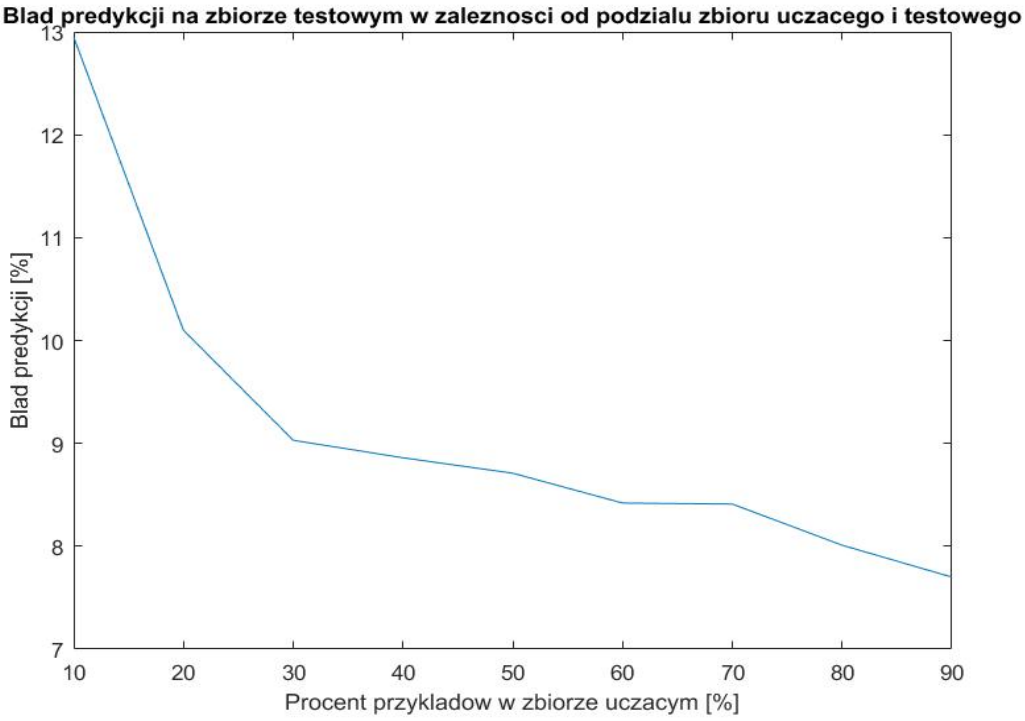
TP = 178	FN = 43
FP = 41	TN = 789

Błąd $\approx 8.01\%$

- 90 %

TP = 87	FN = 21
FP = 19	TN = 398

Błąd $\approx 7.70\%$



Rysunek 3:

4.2.2 Rozcięcia typu akceptorowego

- 10 %

TP = 515	FN = 490
FP = 543	TN = 3661

Błąd $\approx$ 19.84%

- 20 %

TP = 468	FN = 420
FP = 458	TN = 3284

Błąd $\approx$ 18.97%

- 30 %

TP = 428	FN = 348
FP = 380	TN = 2894

Błąd $\approx$ 18.00%

- 40 %

TP = 383	FN = 285
FP = 311	TN = 2492

Błąd $\approx 17.19\%$

- 50 %

TP = 324	FN = 234
FP = 265	TN = 2070

Błąd $\approx 17.27\%$

- 60 %

TP = 260	FN = 182
FP = 210	TN = 1663

Błąd $\approx 16.96\%$

- 70 %

TP = 192	FN = 140
FP = 149	TN = 1255

Błąd $\approx 16.68\%$

- 80 %

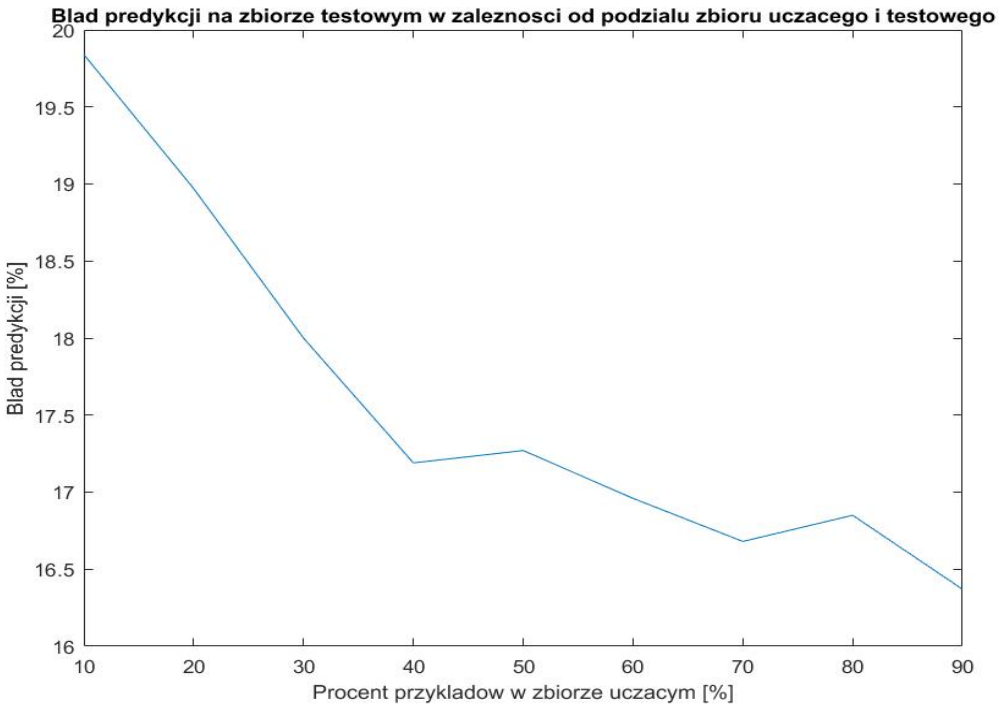
TP = 132	FN = 95
FP = 99	TN = 830

Błąd $\approx 16.85\%$

- 90 %

TP = 63	FN = 44
FP = 49	TN = 420

Błąd $\approx 16.37\%$



Rysunek 4:

W obu przypadkach zauważamy że następuje wraz ze wzrostem ilości przykładów w zbiorze uczącym, spadek błędu predykcji. Jest to zgodne z oczekiwaniami mówiącymi że im większa liczba przykładów w zbiorze uczącym, tym lepiej algorytm może nauczyć się rozdzielać klasy. Porównując algorytm dla rozcięć typu donorowego i akceptorowego zauważamy, że ze względu na podobną liczbę przykładów (odpowiednio 5256 i 5788), różnica w zakresie błędów wynika najprawdopodbniej z długości ciągów znakowych (odpowiednio 15 i 90). Jest to również zgodne z oczekiwaniem, że dla bardziej skomplikowanych przykładów prawdopodobieństwo popełnienia pomyłki przy klasyfikacji jest większe.

4.3 Eksperyment 2.

4.3.1 Rozcięcia typu donorowego

- 10 %
  - Dla zbioru walidującego

TP = 461	FN = 236
FP = 262	TN = 2351

Błąd $\approx$ 15.07%

  - Dla zbioru testowego

TP = 235	FN = 105
FP = 117	TN = 1118

Błąd $\approx 14.12\%$

- 20 %
  - Dla zbioru walidującego

TP = 456	FN = 167
FP = 165	TN = 2155

Błąd $\approx 11.29\%$

  - Dla zbioru testowego

TP = 245	FN = 92
FP = 78	TN = 1160

Błąd $\approx 10.83\%$

- 30 %
  - Dla zbioru walidującego

TP = 420	FN = 129
FP = 128	TN = 1897

Błąd $\approx 10.02\%$

  - Dla zbioru testowego

TP = 267	FN = 69
FP = 82	TN = 1157

Błąd $\approx 9.63\%$

- 40 %
  - Dla zbioru walidującego

TP = 351	FN = 109
FP = 105	TN = 1640

Błąd $\approx 9.76\%$

  - Dla zbioru testowego

TP = 271	FN = 66
FP = 71	TN = 1167

Błąd $\approx 8.74\%$

- 50 %
  - Dla zbioru walidującego

TP = 312	FN = 88
FP = 82	TN = 1356

Błąd $\approx 9.28\%$

- Dla zbioru testowego

TP = 265	FN = 66
FP = 73	TN = 1171

Błąd $\approx 8.88\%$

- 60 %
  - Dla zbioru walidującego

TP = 234	FN = 69
FP = 64	TN = 1103

Błąd $\approx 9.08\%$

  - Dla zbioru testowego

TP = 269	FN = 63
FP = 63	TN = 1180

Błąd $\approx 8.09\%$

- 70 %
  - Dla zbioru walidującego

TP = 188	FN = 49
FP = 48	TN = 818

Błąd $\approx 8.80\%$

  - Dla zbioru testowego

TP = 269	FN = 68
FP = 73	TN = 1165

Błąd $\approx 9.03\%$

- 80 %
  - Dla zbioru walidującego

TP = 124	FN = 34
FP = 30	TN = 547

Błąd $\approx 8.80\%$

  - Dla zbioru testowego

TP = 270	FN = 62
FP = 67	TN = 1176

Błąd $\approx 8.24\%$

- 90 %
  - Dla zbioru walidującego

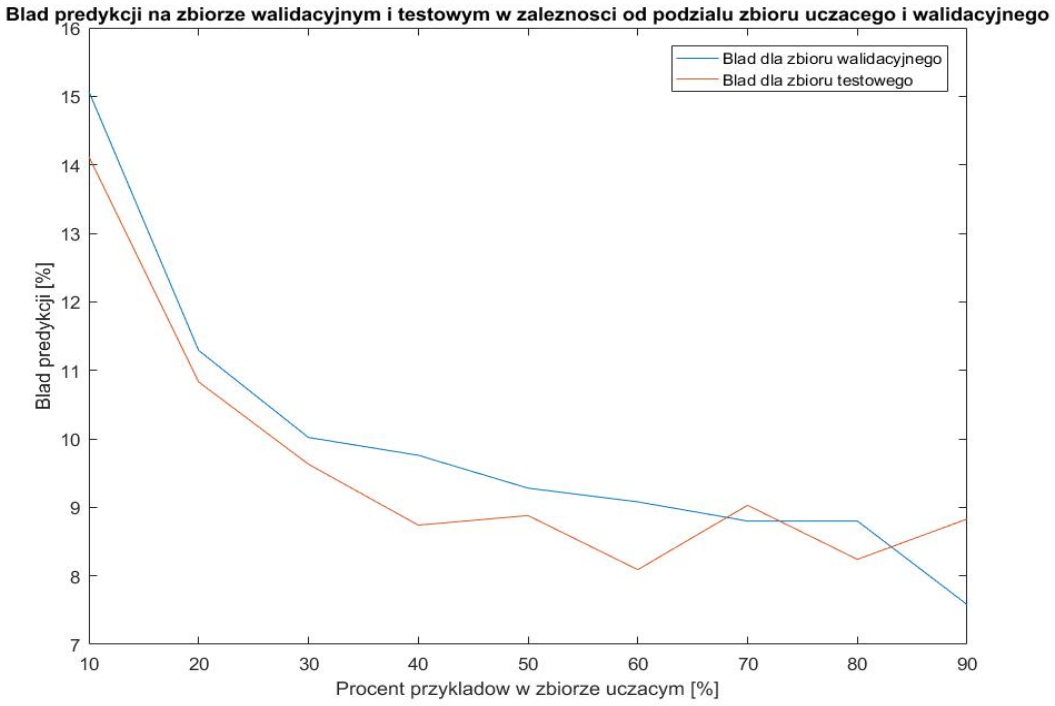
TP = 62	FN = 14
FP = 13	TN = 277

Błąd $\approx 7.58\%$

  - Dla zbioru testowego

TP = 269	FN = 78
FP = 65	TN = 1165

Błąd $\approx 8.83\%$



Rysunek 5:

4.3.2 Rozcięcia typu akceptorowego

- 10 %
  - Dla zbioru walidującego

TP = 349	FN = 352
FP = 404	TN = 2540

Błąd $\approx 20.75\%$

  - Dla zbioru testowego

TP = 156	FN = 178
FP = 159	TN = 1243

Błąd $\approx 19.42\%$

- 20 %
  - Dla zbioru walidującego

TP = 318	FN = 308
FP = 326	TN = 2287

Błąd $\approx 19.59\%$

  - Dla zbioru testowego

TP = 165	FN = 166
FP = 155	TN = 1249

Błąd $\approx 18.53\%$

- 30 %
  - Dla zbioru walidującego

TP = 298	FN = 246
FP = 273	TN = 2017

Błąd $\approx 18.33\%$

  - Dla zbioru testowego

TP = 181	FN = 159
FP = 150	TN = 1246

Błąd $\approx 17.81\%$

- 40 %
  - Dla zbioru walidującego

TP = 259	FN = 210
FP = 235	TN = 1725

Błąd $\approx 18.33\%$

- Dla zbioru testowego

TP = 191	FN = 141
FP = 146	TN = 1257

Błąd $\approx 16.57\%$

- 50 %
  - Dla zbioru walidującego

TP = 228	FN = 166
FP = 183	TN = 1446

Błąd $\approx 17.29\%$

  - Dla zbioru testowego

TP = 198	FN = 134
FP = 154	TN = 1249

Błąd $\approx 16.62\%$

- 60 %
  - Dla zbioru walidującego

TP = 186	FN = 131
FP = 141	TN = 1161

Błąd $\approx 16.83\%$

  - Dla zbioru testowego

TP = 201	FN = 137
FP = 150	TN = 1248

Błąd $\approx 16.56\%$

- 70 %
  - Dla zbioru walidującego

TP = 135	FN = 100
FP = 110	TN = 869

Błąd $\approx 17.34\%$

  - Dla zbioru testowego

TP = 196	FN = 137
FP = 154	TN = 1248

Błąd $\approx 16.83\%$

- 80 %
  - Dla zbioru walidującego

TP = 88	FN = 66
FP = 70	TN = 585

Błąd $\approx 16.97\%$

  - Dla zbioru testowego

TP = 188	FN = 139
FP = 159	TN = 1248

Błąd $\approx 17.24\%$

- 90 %
  - Dla zbioru walidującego

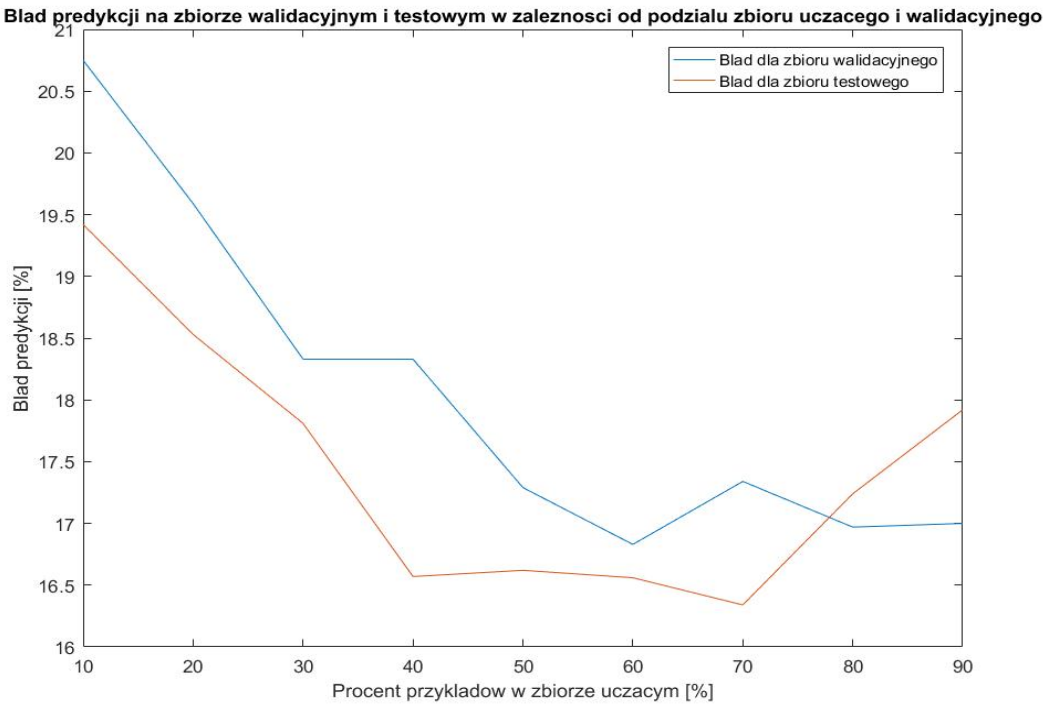
TP = 44	FN = 36
FP = 36	TN = 288

Błąd $\approx 17.86\%$

  - Dla zbioru testowego

TP = 190	FN = 146
FP = 164	TN = 1235

Błąd $\approx 17.92\%$



Rysunek 6:

Zauważamy że wraz ze wzrostem ilości przykładów trenujących, spada błąd predykcji zarówno na zbiorze walidacyjnym (przed przycięciem) jak i na zbiorze testowym (po przycięciu). Przycięcie drzewa powoduje w obu przypadkach poprawę jakości predykcji, dla liczby przykładów przeznaczonej do uczenia mniejszej od ok 60 % - 70 % rozmiaru zbioru uczenie + walidacja. Wynika to najprawdopodobniej z tego że dla dużej liczby przykładów uczących, zbudowane drzewo zbliża się do swojej granicznej możliwości predykcji, a wycinanie niewiele wnosi lub wręcz pogarsza przewidywanie. Porównując algorytm dla rozcięć donorowych jaki i akceptorowych, podobnie jak w poprzednim eksperymencie zauważamy że dla bardziej skomplikowanego zbioru danych, błędy predykcji są większe.

5 Nabyte zdolności

- Zapoznanie się z teorią i praktyką dotyczącą budowy, obsługi i przycinania drzew decyzyjnych.
- Zapoznanie się z rodzajami rozcięć w nici DNA
- Zapoznanie się z referencjami na wskaźniki oraz funkcjami wbudowanymi w niektórych bibliotekach
- Zapoznanie się z możliwością budowy drzew w pakiecie TikZ języka LaTeX

6 Bibliografia

[1] Wykłady do przedmiotu Uczenie się Maszyn

[2] P.Cichosz *Systemy Uczące się*, Wydanie Drugie, Wydawnictwo Naukowo-Techniczne, Warszawa 2000, 2007

[3] https://en.wikipedia.org/wiki/Confusion_matrix