Sprawozdanie z projektu
UMA

Krystian Czechowicz, 318369
Bartosz Niemiec, 318393

Treść zadania

Drzewo decyzyjne w zadaniu klasyfikacji miejsc rozcięcia w sekwencji DNA. Należy dopuścić alternatywę w testach, np. if(atr12=='A'||atr12=='T'). Więcej informacji o specyfice problemu znaleźć można w: https://staff.elka.pw.edu.pl/~rbiedrzy/UMA/opisDNA.html. Dane do pobrania: donory: https://staff.elka.pw.edu.pl/~rbiedrzy/UMA/spliceDTrainKIS.dat, akceptory: https://staff.elka.pw.edu.pl/~rbiedrzy/UMA/spliceATrainKIS.dat. Przed rozpoczęciem realizacji projektu proszę zapoznać się z zawartością: https://staff.elka.pw.edu.pl/~rbiedrzy/UMA/index.html.

Interpretacja zadania
Zadaniem implementowanego przez nas klasyfikatora jest rozpoznawanie miejsc rozcięcia w sekwencji DNA. Jego wystąpienie jest sygnalizowane pojawieniem się w ciągu pewnej specjalnej sekwencji nukleotydów, lecz nie zawsze gwarantuje to napotkanie faktycznego miejsca rozcięcia. Z tego powodu dane trenujące zawierają przykłady pozytywne i takie które wyglądają jakby były pozytywne, lecz nimi nie są. Trening przeprowadzony zostanie osobno dla rozpoznawania donorów i osobno dlaakceptorów. Dane wejściowe rozdzielone są przez to na 2 pliki. Po zbudowaniu drzewa decyzyjnego algorytm powinien z satysfakcjonującą dokładnością rozpoznawać miejsca rozcięcia, bazując na tym jaką sekwencję DNA dostał na wejściu. Atrybutami klasyfikatora będą kolejne miejsca w sekwencji. W węzłach wewnętrznych sprawdzane będą warunki przejścia w głąb drzewa zapewniające możliwość występowania alternatywy (wystąpienie C lub G na 5 miejscu). Przejście może skierować nas do liścia (klasa przyjmuje wartość 0 - nie występuje miejsce rozcięcia, lub 1 - występuje miejsce rozcięcia) lub do kolejnego węzła wewnętrznego (jeśli nie osiągnęliśmy węzła terminalnego). Klasyfikacja na zbiorze testującym powinna przebiec jak najefektywniej (testowanie różnych kryteriów podziału i stopu) i jak najszybciej (implementacja przycinania).

Uruchomienie programu

W celu uruchomienia programu należy wykonać następujące komendy:
> cd {katalog projektu}
> python3 ./main.py [nazwa piku danych] [kryterium podziału] [kryterium stopu] [uwzględnij alternatywę] [głębokość drzewa] [procent dominacji klasy] [procent zbioru trenującego]

Opcje jakie należy podać to odpowiednio:
- **nazwa pliku danych** (np. donors.txt)
- **kryterium podziału** - należy wybrać jedną z 2 możliwości:

0 - warunkowy indeks Giniego
1 - entropia warunkowa

- **kryterium stopu** - należy wybrać jedną z 4 możliwości:

0 - podstawowe kryterium stopu
1 - rozluźnione kryterium stopu
2 - podstawowe kryterium stopu + warunek dodatkowy: maksymalna głębokość
3 - rozluźnione kryterium stopu + warunek dodatkowy: maksymalna głębokość

- **uwzględnij alternatywę**:

0 - brak możliwości alternatywy
1 - istnieje możliwość alternatywy

- **głębokość** - maksymalna głębokość drzewa (jedynie w przypadku podania kryterium stopu wynoszącego 2 lub 3)

- **procent dominacji klasy** - jaki procent musi stanowić najliczniejsza klasa, aby drzewo zakończyło się budować (jedynie w przypadku kryterium stopu 1 lub 3)

- **procent zawartości zbioru trenującego** - jaki procent pierwotnego zbioru trenującego zatrzyma budowę drzewa (jedynie w przypadku kryterium stopu 1 lub 3)

Przykładowe wywołanie programu *python3 ./main.py donors.txt 1 0 0* wywoła budowę drzewa decyzyjnego, na bazie danych z pliku donors.txt, entropia warunkowa, kryterium stopu podstawowe, bez możliwości alternatywy w testach (głębokość została pominięta ponieważ kryterium stopu jej nie uwzględnia)

Ocena jakości algorytmu / testy

Ogólne warunki testów:
Wszystkie testy zostały przeprowadzone przy użyciu walidacji krzyżowej.
Kolejne kroki postępowań wyglądały następująco:

1. podzielono zbiór danych na pięć podzbiorów zachowując proporcję klas zawartych w danych oryginalnych
2. nauka drzewa została wykonana pięciokrotnie, za każdym razem z pominięciem jednego z podzbiorów - zbiór ten posłużył jako zbiór walidujący dane drzewo decyzyjne
3. wybrano najlepsze drzewo na bazie dokładności na zbiorze walidacyjnym

Zmiana kolejności przykładów w obrębie podzbiorów nie jest konieczna, gdyż algorytm budowy drzewa wykonuje operacje na całym zbiorze danych trenujących.

Wykorzystane wskaźniki jakości:

Dokładność: stosunek poprawnie zaklasyfikowanych przykładów do wszystkich przykładów. W naszym kontekście powinna wynosić powyżej 80%, ponieważ w takiej proporcji rozłożone są klasy w zbiorach. Wartość 80% można uzyskać drzewem generującym jedynie predykcje w postaci klasy negatywnej.

Liczba prawdziwie pozytywnych (tp): Liczba przykładów prawidłowo zaklasyfikowanych jako pozytywne

Liczba prawdziwie negatywnych (tn): Liczba przykładów prawidłowo zaklasyfikowanych jako negatywne

Liczba fałszywie pozytywnych (fp): Liczba przykładów nieprawidłowo zaklasyfikowanych jako pozytywne

Liczba fałszywie negatywnych (fn): Liczba przykładów nieprawidłowo zaklasyfikowanych jako negatywne

Stosunek prawdziwie pozytywnych do wszystkich pozytywnych (recall):

$$recall = \frac{tp}{tp+fn}$$

Mówi o zdolności klasyfikatora do rozpoznawania przypadków pozytywnych, nie mówi jednak ile przypadków pozytywnych zostało rozpoznanych jako negatywny i ominiętych

Stosunek prawdziwie negatywnych do wszystkich negatywnych (specificity):

$$specificity = \frac{tn}{tn+fp}$$

Stosunek prawdziwie pozytywnych do wszystkich zaklasyfikowanych jako pozytywne (precision):

$$precision = \frac{tp}{tp+fp}$$

Informuje o tym jak dużo przypadków pozytywnych zostało źle sklasyfikowanych lecz traci informacje o tym jak dużo przypadków negatywnych zostało sklasyfikowanych negatywnie

F1 score:
Procent dokładnie przewidzianych przykładów testowych może być niezawodnym wskaźnikiem jakości tylko w przypadku równego zbalansowania danych, tzn każda z klas ma wśród danych mniej więcej tyle samo próbek. Z uwagi na to, że nasz zbiór zawiera dominację klasy “0”, odpowiednim do zastosowania wskaźnikiem będzie F1 Score, łączący w sobie precision i recall, a dokładnie jest ich średnia harmoniczna. Na wynik wpływają nie tylko wartości wskaźników lecz również to jak bardzo się od siebie różnią. Im bardziej tym gorszy wynik otrzymamy z F1 score.

$$F1score = \frac{2}{\frac{1}{precision}+\frac{1}{recall}}$$

Test kryterium podziału:
W trakcie przeprowadzania testów kryterium podziału, wykorzystano podstawowe kryterium stopu oraz opcję bez alternatywy.

- warunkowy indeks Giniego

Zbiory	donory		akceptory	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	91,44%	91,06%	89,33%	83,77%
Prawdziwie pozytywne	665	155	644	135
Prawdziwie negatywne	3190	802	3492	835
Fałszywie pozytywne	122	26	245	100
Fałszywie negatywne	238	68	249	88
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	73,35%	69,51%	72,12%	60,54%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	96,32%	96,86%	93,44%	89,30%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	84,30%	85,64%	72,44%	57,44%
F1 score	78,44%	76,73%	72,28%	58,95%

- entropia warunkowa

Zbiory	donory		akceptory	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	91,65%	90,87%	88,06%	84,44%
Prawdziwie pozytywne	717	181	615	132
Prawdziwie negatywne	3136	775	3463	845
Fałszywie pozytywne	176	53	275	89
Fałszywie negatywne	175	43	278	91
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	80,38%	80,80%	68,87%	59,19%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	94,67%	93,60%	92,64%	90,47%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	80,29%	77,35%	69,10%	59,46%
F1 score	80,34%	79,04%	68,98%	59,46%

Wnioski - porównanie kryteriów podziału:

Eksperymenty dla drzewa z różnymi kryteriami podziału dały zbliżone rezultaty. Oba wyniki cechują się wysoką skutecznością (na poziomie 90%), jednak tylko pozornie jest to zadowalający rezultat. Stosunek prawdziwie pozytywnych do wszystkich pozytywnych próbek, szczególnie w akceptorach, świadczy o tendencji do przewidywania, że próbka nie jest miejscem rozcięcia. Wysokie niezbalansowanie danych na korzyść klas negatywnych pozwala modelowi na osiągnięcie wysokiej skuteczności, lecz klasy pozytywne są przewidywane z dużo niższą skutecznością. Na tym etapie nie można stwierdzić, że jedno z kryteriów podziału ma wyraźną przewagę nad drugim.

Test kryterium stopu:
W trakcie przeprowadzania testów kryterium stopu, kryterium podziału było ustawione na entropię warunkową, nie uwzględniono alternatywy.

Na początku należy znaleźć najlepsze parametry dla rozluźnionego kryterium stopu:

- dominacja klasy (jaki procent powinna stanowić najliczniejsza klasa aby zatrzymać budowę drzewa
- zbyt mała liczba przykładów (jaki procent początkowej liczby przykładów powinien zatrzymać budowę drzewa)

dostrajanie dominacji klasy

Zbiór	donory							
Dominacja	99%		95%		90%		85%	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	91,70%	91,44%	91,94%	91,54%	91,22%	91,44%	91,12%	92,10%
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	80,38%	80,80%	80,04%	80,36%	75,78%	78,13%	73,91%	73,99%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	94,74%	94,32%	95,14%	94,57%	95,38%	95,04%	95,77%	96,98%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	80,47%	79,38%	81,60%	80,00%	81,02%	81,02%	82,50%	86,84%
F1 score	80,43%	80,09%	80,81%%	80,17%	79,56%	79,55%	77,97%	79,90%

Zbiór	akceptory							
Dominacja	99%		95%		90%		85%	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	88,21%	84,53%	88,96%	85,74%	89,89%	86,34%	90,09%	86,78%
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	68,76%	59,19%	66,85%	58,29%	67,41%	57,85%	65,85%	56,95%

Stosunek prawdziwie negatywnych do wszystkich negatywnych	92,85%	90,58%	94,25%	92,29%	95,26%	93,15%	95,88%	93,95%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	69,69%	60,00%	73,52%	64,36%	77,28%	66,84%	79,25%	69,02%
F1 score	69,22%	59,59%	70,03%	61,18%	72,01%	62,02%	71,93%	62,41%

dostrajanie zbyt małej liczba przykładów

Zbiór	donory							
Zawartość zbioru	0.05%		1%		2%		5%	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	90,58%	91,63%	90,27%	91,25%	88,72%	90,58%	87,49%	90,10%
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	73,46%	71,74%	73,65%	75,89%	71,55%	73,54%	63,05%	67,71%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	95,20%	96,98%	94,75%	95,41%	93,36%	95,16%	94,08%	96,71%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	80,49%	86,48%	79,06%	81,73%	74,39%	80,39%	74,18%	82,51%
F1 score	76,81%	78,43%	76,26%	78,70%	72,95%	76,81%	68,16%	74,38%

Zbiór	akceptory							
Zawartość zbioru	0,05%		1%		2%		5%	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	86,80%	87,13%	85,83%	86,10%	85,83%	86,10%	80,71%	80,74%
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	68,75%	71,30%	68,76%	69,96%	68,76%	69,96%	0%	0%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	91,11%	90,91%	89,91%	89,95%	89,91%	89,95%	100%	100%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	64,90%	65,16%	61,96%	62,40%	61,96%	62,40%	-	-
F1 score	66,78%	68,09%	65,18%	65,96%	65,18%	65,96%	-	-

Wnioski - najlepsze ustawienia rozluźnionego kryterium stopu

Na początek skupiając się na kryterium dominacji klasy, można dostrzec, że ogólna skuteczność poprawia się wraz ze spadkiem procentu dominacji poszczególnej klasy. Niestety pomimo poprawy ogólnej jakości predykcji, znacząco traci na tym zdolność do poprawnej klasyfikacji sytuacji wystąpienia miejsca rozcięcia. Predykcja braku miejsca rozcięcia ma natomiast świetną skuteczność sięgającą prawie 97% w przypadku donorów i 94% w przypadku akceptorów. Należy zatem zastanowić się czy takie rezultaty nas zadowalają. Jeżeli drzewo decyzyjne miałoby posłużyć za wstępną selekcję danych i odrzucanie przypadków, w których miejsca rozcięcia nie występują, tak wytrenowane drzewo sprawdziłoby się bardzo dobrze. Jeżeli jednak zależy nam na tym, aby fragment DNA był zaklasyfikowany jako miejsce rozcięcia z bardzo dobrą skutecznością (bez konieczności dalszego badania zbioru danych zaklasyfikowanych jako pozytywne), należy rozważyć inny parametr rozluźnionego kryterium stopu.

Można zauważyć tendencję do spadku dokładności klasyfikacji przypadków pozytywnych wraz ze spadkiem wymaganego procentu dominacji jednej z klas. Wynika to najprawdopodobniej z faktu, że w początkowym zbiorze istnieje znaczna dysproporcja w liczbie klas. Szybsze zakończenie budowy drzewa, skutkuje brakiem możliwości ustalenia kiedy przypadek klasyfikować jako pozytywne miejsce rozcięcia, natomiast drzewo zaczyna specjalizować się w rozpoznawaniu przypadków negatywnych.

Na potrzeby dalszych eksperymentów uznamy, że najlepiej sprawdza się dominacja klasy na poziomie 85%, gdyż daje ona najbardziej różne od wariantu podstawowego wyniki. Będziemy zatem w stanie porównać jakość drzewa decyzyjnego na różnych płaszczyznach.

Inaczej sytuacja wygląda z minimalną wielkością zbioru trenującego. W przypadku danych akceptorów, wprowadzenie takie wariantu rozluźnionego poprawia jakość drzewa decyzyjnego w szczególności w kontekście klasyfikacji klas pozytywnych. Nie należy jednak przesadzić z wielkością tego parametru, gdyż może to skutkować modelem, który każdy przypadek klasyfikuje jako jedna klasa. Jednak w przypadku drugiego ze zbiorów danych jakość ulega pogorszeniu. Należy zatem zauważyć, że dobór parametrów wariantu rozluźnionego jest zagadnieniem złożonym i może zależeć od specyfiki zbioru danych.

Można zatem uznać, że najlepszym wyborem będzie niewielka minimalna wielkość zbioru trenującego (w naszym przypadku 0,05% pierwotnej wielkości zbioru)

Finalne parametry wyglądają zatem następująco:
Aby zatrzymać budowę drzewa najliczniejsza klasa musi stanowić 85% wszystkich klas oraz minimalna wielkość zbioru wynosi 0,05% oryginalnej wielkości zbioru.

Porównanie podstawowego kryterium stopu i dostrojonego kryterium stopu

Zbiór	donory				akceptory			
kryterium	podstawowe		rozluźnione		podstawowe		rozluźnione	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	91,65%	90,87%	90,39%	91,16%	88,06%	84,44%	87,80%	87,91%
Prawdziwie pozytywne	717	181	616	161	615	132	595	158
Prawdziwie negatywne	3136	775	3184	798	3463	845	3470	860
Fałszywie pozytywne	176	53	128	30	275	89	267	75
Fałszywie negatywne	175	43	276	63	278	91	298	65
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	80,38%	80,80%	69,06%	71,88%	69,87%	59,19%	66,63%	70,85%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	94,67%	93,60%	96,14%	96,38%	92,64%	90,47%	92,86%	91,98%

Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	80,29%	77,35%	82,80%	84,29%	69,10%	59,73%	69,03%	67,81%
F1 score	80,34%	79,03%	75,31%	77,59%	68,98%	59,46%	67,81%	69,30%

Wnioski - porównanie podstawowego i rozluźnionego kryterium stopu

Jak można zauważyć, jakość drzewa decyzyjnego zależy nie tylko od dobranych parametrów, ale także od samego zbioru danych. Trzeba mieć to na uwadze, jeżeli decydujemy się na wykorzystanie drzewa decyzyjnego, którego parametry były ustawiane na bazie innego zbioru. Oczywiście, można próbować wyciągać średnie wartości parametrów na różnych zbiorach, jednak jeżeli zależy nam na tym, aby predykcja klas dla konkretnego przypadku była jak najlepsza należy rozważyć ręczny dobór parametrów. Może okazać się, jak w przypadku zbioru donorów, że próba znalezienia parametrów poprawiających jakość algorytmu zakończy się niepowodzeniem, jednak zdarzyć się może również, że uzyskamy znacznie lepsze wyniki, jak np. w przypadku zbioru akceptorów.

Dodatkowe kryterium stopu - dostrajanie

Test dodatkowego kryterium stopu - maksymalnej głębokości zostanie przeprowadzony zarówno w połączeniu z podstawowym kryterium stopu jak i rozluźnionym kryterium stopu.

Zbiór	donory							
Maksymalna głębokość	brak (kryterium podstawowe)		3		5		7	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	91,65%	90,87%	89,23%	91,91%	91,75%	91,81%	91,46%	91,86%
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	80,38%	80,80%	67,08%	71,74%	83,20%	79,37%	80,49%	81,70%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	94,67%	93,60%	95,20%	97,34%	94,05%	95,17%	94,41%	93,60%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	80,29%	77,35%	79,02%	87,91%	79,04%	81,57%	79,51%	77,54%
F1 score	80,34%	79,03%	72,56%	79,01%	81,07%	80,45%	80,00%	79,57%

Zbiór	akceptory							
Maksymalna głębokość	brak (kryterium podstawowe)		3		5		6	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	88,06%	84,44%	86,46%	87,56%	88,34%	85,05%	88,04%	84,96%
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	68,87%	59,19%	72,00%	77,58%	74,47%	64,13%	70,10%	62,33%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	92,64%	90,47%	89,91%	89,95%	91,65%	90,04%	92,32%	90,36%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	69,10%	59,73%	63,04%	64,79%	68,07%	60,59%	68,57%	60,70%
F1 score	68,98%	59,46%	67,22%	70,61%	71,12%	62,31%	69,32%	61,50%

Zbiór	donory							
Maksymalna głębokość	brak (kryterium rozluźnione)		2		3		4	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	90,39%	91,16%	87,49%	90,10%	89,01%	91,44%	90,39%	91,16%
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	69,06%	71,88%	63,04%	67,71%	62,93%	69,06%	69,06%	71,88%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	96,14%	96,38%	94,08%	96,14%	96,04%	97,46%	96,14%	96,38%
Stosunek prawdziwie pozytywnych do przewidywanych	82,80%	84,29%	74,18%	82,51%	81,10%	88,00%	82,80%	84,29%

jako pozytywne								
F1 score	75,31%	77,59%	68,16%	74,38%	70,87%	77,39%	75,30%	77,59%

Zbiór	akceptory							
Maksymalna głębokość	brak (kryterium rozluźnione)		2		3		4	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	87,80%	91,86%	81,79%	82,31%	86,46%	87,56%	87,80%	87,91%
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	66,63%	70,85%	19,39%	18,75%	72,00%	77,58%	66,63%	70,85%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	92,86%	91,98%	96,68%	97,54%	89,91%	89,94%	92,86%	91,98%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	69,03%	67,81%	58,24%	64,62%	63,04%	64,79%	69.03%	67,81%
F1 score	67,81%	69,30%	29,10%	29,07%	67,22%	70,61%	67,81%	69,30%

Wnioski - kryterium dodatkowe: głębokość

Podstawowym wnioskiem płynącym z powyższych eksperymentów jest fakt, że dobór wartości skrajnych, atrybutu dodatkowego w postaci maksymalnej głębokości drzewa, nie jest dobrym pomysłem. Jeżeli atrybut jest zbyt duży, nie będzie mieć żadnego wpływu na działanie algorytmu (inne warunki stopu zatrzymają budowę szybciej). Jeżeli będzie zbyt mały budowa drzewa zatrzyma się przedwcześnie co znacząco wpłynie na jego jakość. Jednakże w każdym z powyższych przykładów udało się znaleźć taką wartość maksymalnej głębokości, która poprawia ogólne działanie algorytmu. Niestety jaką dokładnie wartość powinien przyjąć ten atrybut, zależy w dużej mierze od pozostałych atrybutów algorytmu oraz od zbioru danych. Tendencja, którą można się kierować to dopuszczenie głębszego drzewa jest lepsze, aż do pewnej wartości granicznej.

Wnioski - porównanie wszystkich kryteriów stopu

Zgodnie z powyższymi wnioskami kryteria stopu powinno dobrać się odpowiednio do konkretnego zbioru danych. Najlepsze okazują się zatem:

Zbiór donorów: kryterium podstawowe + maksymalna głębokość 7
Zbiór akceptorów: kryterium rozluźnione/kryterium podstawowe + maksymalna głębokość 3

W przypadku zbioru akceptorów kluczowym atrybutem dla jego jakości okazała się głębokość niezależnie od wariantu kryterium stopu.

Test uwzględnienia alternatywy:
Do testów alternatywy posłużymy się podstawowym kryterium stopu oraz entropią warunkową do wyznaczenia podziału.

Zbiór	donory				akceptory			
alternatywa	brak		uwzględnij		brak		uwzględnij	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	91,65%	90,87%	91,65%	90,87%	88,06%	84,44%	88,06%	84,44%
Prawdziwie pozytywne	717	181	717	181	615	132	615	132
Prawdziwie negatywne	3136	775	3136	775	3463	845	3463	845
Fałszywie pozytywne	176	53	176	53	275	89	275	89
Fałszywie negatywne	175	43	175	43	278	91	278	91
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	80,38%	80,80%	80,38%	80,80%	69,87%	59,19%	69,87%	59,19%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	94,67%	93,60%	94,67%	93,60%	92,64%	90,47%	92,64%	90,47%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	80,29%	77,35%	80,29%	77,35%	69,10%	59,73%	69,10%	59,73%
F1 score	80,34%	79,03%	80,34%	79,03%	68,98%	59,46%	68,98%	59,46%

Test alternatywy - niepokojące wyniki:

Według powyższej tabeli jakość algorytmu jest dokładnie taka. Wynik ten może wydawać się niepokojący, nieprawidłowy. Należy zatem zagłębić się w działanie algorytmu i znaleźć przyczynę takiego stanu rzeczy.

Po przeprowadzeniu inspekcji działania algorytmu okazuje się, że sposób podziału jaki jest wybierany, zawsze odpowiada podziałowi na największą możliwą liczbę gałęzi (podział na 4 poddrzewa z testem na nukleotydy bez alternatywy). Podejrzana o taki stan powinna być zatem entropia warunkowa. Przeprowadźmy zatem analogiczny eksperyment, tym razem nie dopuśćmy jednak do możliwości wyboru opcji bez alternatywy (podział na 4 poddrzewa jest niemożliwy).

Rezultaty wyglądają następująco:

Zbiór	donory				akceptory			
alternatywa	brak		uwzględnij		brak		uwzględnij	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	91,65%	90,87%	90,89%	91,53%	88,06%	84,44%	85,40%	84,47%
Prawdziwie pozytywne	717	181	652	159	615	132	659	153
Prawdziwie negatywne	3136	775	3170	803	3463	845	3294	826
Fałszywie pozytywne	176	53	142	25	275	89	443	109
Fałszywie negatywne	175	43	241	64	278	91	223	71
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	80,38%	80,80%	73,01%	71,30%	69,87%	59,19%	73,88%	68,30%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	94,67%	93,60%	95,71%	96,98%	92,64%	90,47%	88,15%	88,34%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	80,29%	77,35%	82,11%	86,41%	69,10%	59,73%	59,80%	58,40%
F1 score	80,34%	79,03%	77,30%	78,13%	68,98%	59,46%	66,10%	62,96%

Jak widać wyniki wyglądają inaczej. Dodatkowo badając wybierane podziały można zauważyć, że zawsze wybierany jest 1 z 6 podziałów na 3 poddrzewa. Przeprowadźmy

zatem jeszcze jeden eksperyment pozostawiając jedynie możliwość podziału na dwa poddrzewa (binarne drzewo decyzyjne).

Rezultaty prezentują się w taki sposób:

Zbiór	donory				akceptory			
alternatywa	brak		uwzględnij		brak		uwzględnij	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	91,65%	90,87%	92,74%	93,05%	88,06%	84,44%	86,33%	85,57%
Prawdziwie pozytywne	717	181	729	188	615	132	444	109
Prawdziwie negatywne	3136	775	3171	790	3463	845	3554	881
Fałszywie pozytywne	176	53	141	38	275	89	184	53
Fałszywie negatywne	175	43	164	35	278	91	449	114
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	80,38%	80,80%	81,63%	84,30%	69,87%	59,19%	49,72%	48,87%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	94,67%	93,60%	95,74%	95,41%	92,64%	90,47%	95,07%	94,32%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	80,29%	77,35%	83,79%	83,19%	69,10%	59,73%	70,60%	67,28%
F1 score	80,34%	79,03%	82,70%	83,74%	68,98%	59,46%	58,38%	56,62%

Dodatkowo sprawdzony został warunkowy indeks Giniego przy pierwotnej koncepcji uwzględnienia alternatywy:

Zbiór	donory				akceptory			
alternatywa	brak		uwzględnij		brak		uwzględnij	
Dane	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące	Dane trenujące	Dane walidujące
Dokładność	91,44%	91,06%	91,44%	91,06%	89,33%	83,77%	89,33%	83,77%
Prawdziwie pozytywne	655	155	655	155	644	135	644	135
Prawdziwie negatywne	3190	802	3190	802	3492	835	3492	835
Fałszywie pozytywne	122	26	122	26	245	100	245	100
Fałszywie negatywne	238	68	238	68	249	88	249	88
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	73,34%	69,51%	73,34%	69,51%	72,11%	60,54%	72,12%	60,54%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	96,31%	96,86%	96,31%	96,86%	93,44%	89,30%	93,44%	89,30%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	84,30%	85,64%	84,30%	85,64%	72,44%	57,44%	72,44%	57,44%
F1 score	78,44%	76,73%	78,44%	76,73%	72,28%	58,95%	72,28%	58,95%

Jak widać ponownie zbiory są takie same, a wybieranym podziałem jest podział na 4 poddrzewa. Oznacza to, że założenia podziału jakie przyjęliśmy na początku wykluczają możliwość alternatywy. Można to jednak zmienić koncepcję na binarne drzewo decyzyjne, co może poprawić wyniki (np. 93% dokładności na zbiorze donorów).

Wnioski - test alternatywy:

Użycie alternatywy w takiej koncepcji jaką założyliśmy na początku mija się z celem, ponieważ zarówno entropia, jak i warunkowy indeks Giniego zawsze zwraca podział na największą liczbę poddrzew (a przynajmniej na używanych przez nas zbiorach). Jeżeli chcemy, aby alternatywa była rzeczywiście uwzględniana należy rozważyć podział binarny:
A | CTG
T | ACG
C | ATG
G | CAT

AC | GT
AG | CT
AT | CG

Wtedy możemy uzyskać lepsze wyniki niż w podziale na 4 poddrzewa (A | T | C | G).

Najlepsze uzyskane drzewo decyzyjne:

Łącząc najlepsze możliwe konfiguracje dla zbioru danych donorów tj:
kryterium podziału: entropia
kryterium stopu: podstawowe + maksymalna głębokość 7
alternatywa: binarne drzewo decyzyjne z uwzględnieniem alternatywy
wyniki prezentują się następująco:

Zbiór	donory	
alternatywa	uwzględnij	
Dane	Dane trenujące	Dane walidujące
Dokładność	93,81%	93,05%
Prawdziwie pozytywne	774	190
Prawdziwie negatywne	3171	788
Fałszywie pozytywne	141	40
Fałszywie negatywne	119	33
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	86,67%	85,20%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	95,74%	95,17%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	84,59%	82,61%
F1 score	85,62%	83,89%

Łącząc najlepsze możliwe konfiguracje dla zbioru danych akceptorów tj:
kryterium podziału: entropia
kryterium stopu: podstawowe + maksymalna głębokość 3
alternatywa: nie uwzględniaj
wyniki prezentują się następująco:

Zbiór	akceptory	
alternatywa	uwzględnij	
Dane	Dane trenujące	Dane walidujące
Dokładność	86,46%	87,56%
Prawdziwie pozytywne	643	173
Prawdziwie negatywne	3360	841
Fałszywie pozytywne	377	94
Fałszywie negatywne	250	50
Stosunek prawdziwie pozytywnych do wszystkich pozytywnych	72,00%	77,58%
Stosunek prawdziwie negatywnych do wszystkich negatywnych	89,91%	89,95%
Stosunek prawdziwie pozytywnych do przewidywanych jako pozytywne	63,04%	64,79%
F1 score	67,22%	70,61%

Wyniki dla drzewa decyzyjnego z pakietu scikit-learn:
Testy przeprowadzone dla obu kryteriów podziału, z podstawowym kryterium stopu.

WARUNKOWY INDEKS GINIEGO

Zbiór	akceptory
Dane	Dane walidujące
Dokładność	97,70%
F1 score	97,70%

Zbiór	donory
Dane	Dane walidujące
Dokładność	95,60%
F1 score	95,50%

ENTROPIA WARUNKOWA

Zbiór	akceptory
Dane	Dane walidujące
Dokładność	95,50%
F1 score	95,60%

Zbiór	donory
Dane	Dane walidujące
Dokładność	93,30%
F1 score	94%

Drzewo decyzyjne zbudowane przy użyciu biblioteki scikitlearn uzyskało dużo lepsze wyniki niż to zbudowane przez nas. Możemy jednak zbliżyć się do tych wartości ograniczając zbiory danych, eliminując jednocześnie niezbalansowanie klas. Po usunięciu próbek negatywnych tak by klasa ta stanowiła 50% (zastosowane w zbiorze donorów oraz akceptorów) wyniki prezentują się następująco:

WARUNKOWY INDEKS GINIEGO

Zbiór	akceptory
Dane	Dane walidujące
Dokładność	85,40%
F1 score	85,40%

Zbiór	donory
Dane	Dane walidujące
Dokładność	89,00%
F1 score	89,00%

ENTROPIA WARUNKOWA

Zbiór	akceptory
Dane	Dane walidujące
Dokładność	81,09%
F1 score	81,09%

Zbiór	donory
Dane	Dane walidujące
Dokładność	93,30%
F1 score	94,30%

Jak możemy zauważyć skuteczność predykcji nie poprawiła się znacząco, w przeciwieństwie do F1 score. Wynika to z faktu, że ten drugi jest wskaźnikiem określającym skuteczność naszego modelu właśnie w sytuacji niezbalansowania danych. W momencie kiedy je zbalansujemy, traci dodatkową informacje którą niósł, i możemy go utożsamiać z klasyczną skutecznością.

Podsumowanie projektu:

Pomimo, iż algorytmy przez nas stworzone osiągnęły gorsze rezultaty niż zaawansowany algorytm z biblioteki scikit-learn, to wyniki uznajemy za zadowalające.
Głównym wnioskiem płynącym z przeprowadzonych eksperymentów jest z pewnością fakt, iż do każdego zbioru danych należy podchodzić indywidualnie, gdyż różne ustawienia algorytmu mogą dawać różne rezultaty.
Projekt nauczył nas, jak zaawansowanym zagadnieniem potrafią być drzewa decyzyjne. Liczba różnych konfiguracji, jakie można zastosować z jednej strony sprawia, że znalezienie ustawień idealnych może być ciężkie, z drugiej strony jednak może stanowić ogromną zaletę algorytmu, ponieważ staje się on uniwersalny (można go dostosować, do każdego zbioru).
Projekt uświadomił nas również jak na wyniki predykcji wpływa niezbalansowanie danych. Przy równym podziale klas w zbiorze danych drzewo decyzyjne ma dużo lepszą skuteczność niż w przypadku gdy obserwujemy dominacje jednej z nich. Dlatego bardzo ważne jest stosowanie bardziej zaawansowanych wskaźników jakości, co pozwoli nam takie problemy wykrywać i eliminować.