Projekt UMA – Kamil Kośnik, Kacper Radzikowski

Repozytorium projektu: https://github.com/FRSH-0109/ML-DT-DNA

Opis oryginalny dostępny na stronie "zapisy":

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 opisie. Dane do pobrania: donory, akceptory. Przed rozpoczęciem realizacji projektu proszę zapoznać się z zawartością strony.

Opis oryginalny dostępny na stronie prowadzącego (link):

Istnieją dwa rodzaje miejsc rozcięcia sekwencji kodującej białko: donory i akceptory. Ich odnalezienie otwiera drogę do znalezienia eksonów, czyli sekwencji kodujących białka.

Należy zaimplementować klasyfikator, następnie przeprowadzić jego trening i testowanie na 2 problemach:

1. szukanie donorów,
2. szukanie akceptorów

Każdy ze zbiorów danych należy rozdzielić na trenujący i testujący lub zastosować walidację krzyżową. Zaimplementowany klasyfikator należy przebadać (wykonać eksperymenty).

Jeżeli chodzi o dane to w tym pliku znajdują się przykłady donorów, a w tym pliku przykłady akceptorów. W pierwszej linii każdego z nich napisano, na której pozycji (licząc litery od lewej strony) we fragmentach sekwencji jest granica pomiędzy intronem a eksonem. Dana ta jest zbędna dla klasyfikatora - może jednak pomóc badaczowi w interpretacji wyników. Dalej w pliku występują parami: linia określająca czy jest to przykład pozytywny (1) czy negatywny (0) oraz sam przykład, czyli sekwencja DNA. Przykłady negatywne to takie, które częściowo wyglądają jak miejsca rozcięcia, ale nimi nie są.

Interpretacja zadania

Celem zadania jest implementacja algorytmu drzewa decyzyjnego w taki sposób, aby finalny rezultat umożliwiał rozpoznawanie miejsc rozcięcia sekwencji kodującej białko. Model drzewa decyzyjnego będzie realizował dwa zadania klasyfikacji – szukanie donorów i akceptorów. Dodatkowo w ramach zadania przeprowadzone zostaną eksperymenty sprawdzające możliwości i właściwości modelu w zależności od jego parametrów - wartości kryteriów stopu: głębokości drzewa oraz minimalnego zbioru do podziału; dodatkowo porównane zostaną dwie wielkości wykorzystywane w algorytmie - entropię oraz współczynnik Gini'ego. Na bazie wyników tych eksperymentów dokonane zostanie porównanie oraz wyciągnięcie wniosków na temat badanego algorytmu uczenia maszynowego.

Opis algorytmu

Algorytm drzewa decyzyjnego realizowany jest w następujący sposób:

1. Inicjujemy algorytm pierwszym węzłem drzewa decyzyjnego nazywanego korzeniem drzewa (z ang. root) zawierającego cały zbiór danych trenujących **T**
2. Znajdujemy dla zbioru danych w węźle atrybut i próg (przy atrybutach, które są opisywane wartościami ciągłymi) wartości najlepiej maksymalizujący przyrost informacji, wyznaczanego następującym wzorem:

$$IG(A) = ASM(S) - \sum_{v \epsilon Wartości(A)} \frac{|S_v|}{|S|} ASM(S_v)$$

   gdzie,

- $IG(A)$ - przyrost informacji dla atrybutu A
- ASM - Atrybut miary wyboru
- **S** - Zbiór przykładów
- $\mathbf{S_v}$ - Zbiór przykładów ze zbioru **S** dla których parametr **A** przyjmuje wartość **v**
- **Wartości(A)** - zbiór wszystkich możliwych wartości parametru **A**
- $|S|$ - rozmiar zbioru **S**
- $|S_v|$ - rozmiar zbioru $\mathbf{S_v}$

3. Dzielimy zbiór **T** na podzbiory dzieląc je na bazie wyznaczonej w kroku 2 pary wartość-próg (w przypadku, gdy atrybut ma wartości ciągłe)
4. Generujemy nowy węzeł decyzyjny zawierający wybrany najlepszy atrybut
5. Tworzymy rekursywnie nowe drzewa decyzyjne wykorzystując podzbiory danych, które stworzyliśmy w kroku 3 do momentu uzyskania jednego z kryteriów stopu: zbiór danych jest czysty – zawiera tylko przykłady jednej klasy, którą przypisujemy jako wartość ostatniego węzła tzw. Liścia drzewa; osiągniemy maksymalną założoną przez nas głębokość drzewa lub zbiór danych w węźle jest mniejszy niż ustalona przez nas wartość graniczna wielkości zbioru do podziału. W dwóch ostatnich przypadkach w zadaniu klasyfikacji, które będziemy realizować za wartość liścia przyjmujemy klasę, która reprezentuje najwięcej przykładów w zbiorze danych

Rolę Atrybutu Miary Wyboru (z ang. Attribute Selection Measure) może realizować: Entropia oraz Współczynnik Gini'ego. Entropię oznaczaną $H(S)$ opisuje następujący wzór:

$$H(S) = -\sum_{k \epsilon K}^{n} p(k) \cdot \log_2\big(p(k)\big)$$

gdzie,

S - Zbiór przykładów

K - Zbiór możliwych do uzyskania przez przykłady ze zbioru **S** klas

k – konkretna klasa ze zbioru **K**

$p(k)$ - prawdopodobieństwo, że wylosowany przykład ze zbioru **S** jest klasy **k** – efektywnie stosunek przykładów o klasie **k** w zbiorze **S** do wszystkich przykładów w zbiorze.

Współczynnik Gini'ego oznaczany $G(S)$ opisuje następujący wzór:

$$G(S) = 1 - \sum_{k \epsilon K}^{n} p^2(k)$$

gdzie,

S - Zbiór przykładów

$p(k)$ - prawdopodobieństwo, że wylosowany przykład ze zbioru **S** jest klasy **k** – efektywnie stosunek przykładów o klasie **k** w zbiorze **S** do wszystkich przykładów w zbiorze.

Przykładowa realizacja algorytmu

Na potrzeby przykładu będziemy rozpatrywać następujący zbiór danych testowych:

Nr.	a_1	a_2	a_3	Klasa
1	0	0	0	0
2	0	1	0	0
3	0	1	1	1
4	1	0	0	1
5	1	0	1	1
6	1	1	1	1

Gdzie a_1, a_2, a_3 to atrybuty binarne

Realizacja z użyciem Entropii jako Współczynnik Miary Atrybutu

0) Wyznaczamy Entropię węzła początkowego

$$H_0 = -\left(\frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right) + \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right)\right) = -(-0{,}39 - 0{,}5283) = 0{,}9183$$

1) Dla atrybutu a_1

$$a_1 = 0 \Rightarrow [1, 2, 3]$$

$$a_1 = 1 \Rightarrow [4, 5, 6]$$

$$H_{1a_1} = -\left(\frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right) + \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right)\right) = -(-0{,}39 - 0{,}5283) = 0{,}9183$$

$$H'_{1a_1} = -(0 \cdot \log_2(0) + 1 \cdot \log_2 \quad 1) = -(-0 - 0) = 0$$

$$IG_{a_1} = 0{,}9183 - \left(\frac{1}{2} \cdot 0{,}9183 + \frac{1}{2} \cdot 0\right) = 0{,}9183 - 0{,}45915 = 0{,}45915$$

Dla atrybutu a_2

$$a_2 = 0 \Rightarrow [1, 4, 5]$$

$$a_2 = 1 \Rightarrow [2, 3, 6]$$

$$H_{1a_2} = -\left(\frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) + \frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right)\right) = -(-0{,}5283 - 0{,}39) = 0{,}9183$$

$$H'_{1a_2} = -\left(\frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) + \frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right)\right) = -(-0{,}5283 - 0{,}39) = 0{,}9183$$

$$IG_{a_2} = 0{,}9183 - \left(\frac{1}{2} \cdot 0{,}9183 + \frac{1}{2} \cdot 0{,}9183\right) = 0{,}9183 - 0{,}9183 = 0$$

Dla atrybutu a_3

$$a_3 = 0 \Rightarrow [1,\ 2,\ 4]$$

$$a_3 = 1 \Rightarrow [3,\ 5,\ 6]$$

$$H_{1a_3} = -\left(\frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right) + \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right)\right) = -(-0{,}39 - 0{,}5283) = 0{,}9183$$

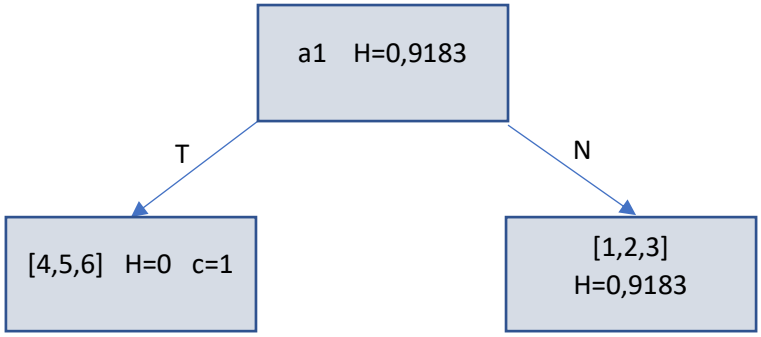
$$H'_{1a_3} = -(0 \cdot \log_2(0) + 1 \cdot \log_2(1)) = -(-0 - 0) = 0$$

$$IG_{a_3} = 0{,}9183 - \left(\frac{1}{2} \cdot 0{,}9183 + \frac{1}{2} \cdot 0\right) = 0{,}9183 - 0{,}45915 = 0{,}45915$$

Dobór atrybutu pierwszego węzła

$IG_{a_1} = IG_{a_3} > IG_{a_2}$ - wybieramy atrybut a_1 z racji na to, że badaliśmy go jako pierwszy – wybranie atrybutu a_3 jest jednak również poprawną selekcją

Drzewo decyzyjne po kroku 1):



2) Z racji na fakt, że atrybuty w naszym przykładzie są binarne atrybut a_1 możemy pominąć w dalszej analizie.

Dla atrybutu a_2

$$a_2 = 0 \Rightarrow [1]$$

$$a_2 = 1 \Rightarrow [2,3]$$

$$H_{2a_2} = -(1 \cdot \log_2(1) + 0 \cdot \log_2(0)) = -(-0 - 0) = 0$$

$$H'_{2a_2} = -\left(\frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right)\right) = -\left(-\frac{1}{2} - \frac{1}{2}\right) = 1$$

$$IG_{a_2} = 0{,}9183 - \left(\frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1\right) = 0{,}9183 - \frac{2}{3} = 0{,}2516$$

Dla atrybutu a_3

$$a_3 = 0 \Rightarrow [1,2]$$

$$a_3 = 1 \Rightarrow [3]$$

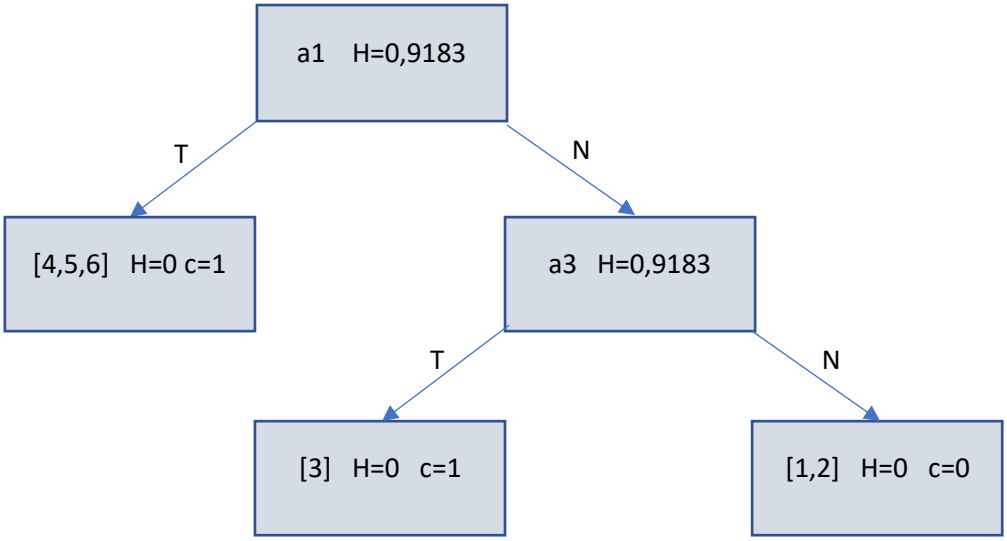
$$H_{2a_3} = -(1 \cdot \log_2(1) + 0 \cdot \log_2(0)) = -(-0 - 0) = 0$$

$$H'_{2a_3} = -(0 \cdot \log_2(0) + 1 \cdot \log_2(1)) = -(-0 - 0) = 0$$

$$IG_{a_3} = 0{,}9183 - \left(\frac{2}{3} \cdot 0 + \frac{1}{3} \cdot 0\right) = 0{,}9183 - 0 = 0{,}9183$$

$IG_{a_3} > IG_{a_2}$ - wybieramy atrybut a_3

Drzewo decyzyjne po kroku 2):



Realizacja z użyciem Współczynnika Gini'ego jako Współczynnik Miary Atrybutu

0) Wyznaczamy Współczynnik Gini'ego węzła początkowego

$$G_0 = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 1 - \left(\frac{4}{9} + \frac{1}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9}$$

1) Dla atrybutu a_1

$$a_1 = 0 \Rightarrow [1, 2, 3]$$

$$a_1 = 1 \Rightarrow [4, 5, 6]$$

$$G_{1a_1} = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 1 - \left(\frac{4}{9} + \frac{1}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9}$$

$$G'_{1a_1} = 1 - [0^2 + 1^2] = 1 - 1 = 0$$

$$IG_{a_1} = \frac{4}{9} - \left(\frac{1}{2} \cdot \frac{4}{9} + \frac{1}{2} \cdot 0\right) = \frac{4}{9} - \left(\frac{2}{9} + 0\right) = \frac{4}{9} - \frac{2}{9} = \frac{2}{9}$$

Dla atrybutu a_2

$$a_2 = 0 \Rightarrow [1, 4, 5]$$

$$a_2 = 1 \Rightarrow [2, 3, 6]$$

$$G_{1a_2} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 1 - \left(\frac{1}{9} + \frac{4}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9}$$

$$G'_{1a_2} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 1 - \left(\frac{1}{9} + \frac{4}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9}$$

$$IG_{a_2} = \frac{4}{9} - \left(\frac{1}{2} \cdot \frac{4}{9} + \frac{1}{2} \cdot \frac{4}{9}\right) = \frac{4}{9} - \left(\frac{2}{9} + \frac{2}{9}\right) = \frac{4}{9} - \frac{4}{9} = 0$$

Dla atrybutu a_3

$$a_3 = 0 \Rightarrow [1, 2, 4]$$

$$a_3 = 1 \Rightarrow [3, 5, 6]$$

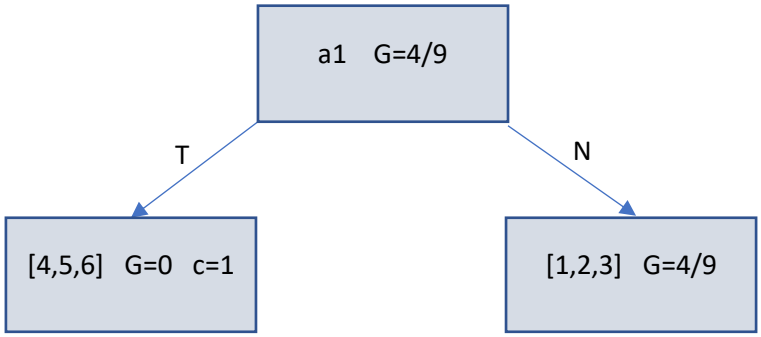
$$G_{1a_3} = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 1 - \left(\frac{4}{9} + \frac{1}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9}$$

$$G'_{1a_3} = 1 - [0^2 + 1^2] = 1 - 1 = 0$$

$$IG_{a_3} = \frac{4}{9} - \left(\frac{1}{2} \cdot \frac{4}{9} + \frac{1}{2} \cdot 0\right) = \frac{4}{9} - \left(\frac{2}{9} + 0\right) = \frac{4}{9} - \frac{2}{9} = \frac{2}{9}$$

$IG_{a_1} = IG_{a_3} > IG_{a_2}$ - wybieramy atrybut z racji na to, że badaliśmy go jako pierwszy wybranie atrybutu jest jednak również poprawną selekcją

Drzewo decyzyjne po kroku 1):



2) Z racji na fakt, że atrybuty w naszym przykładzie są binarne atrybut a_1 możemy pominąć w dalszej analizie.

Dla atrybutu a_2

$$a_2 = 0 \Rightarrow [1]$$

$$a_2 = 1 \Rightarrow [2,3]$$

$$G_{2a_2} = 1 - 1^2 = 1 - 1 = 0$$

$$G'_{2a_2} = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 1 - \left(\frac{1}{4} + \frac{1}{4}\right) = 1 - \frac{1}{2} = \frac{1}{2}$$

$$IG_{a_2} = \frac{4}{9} - \left(\frac{1}{3} \cdot 0 + \frac{2}{3} \cdot \frac{1}{2}\right) = \frac{4}{9} - \frac{1}{3} = \frac{1}{9}$$

Dla atrybutu a_3

$$a_3 = 0 \Rightarrow [1,2]$$

$$a_3 = 1 \Rightarrow [3]$$

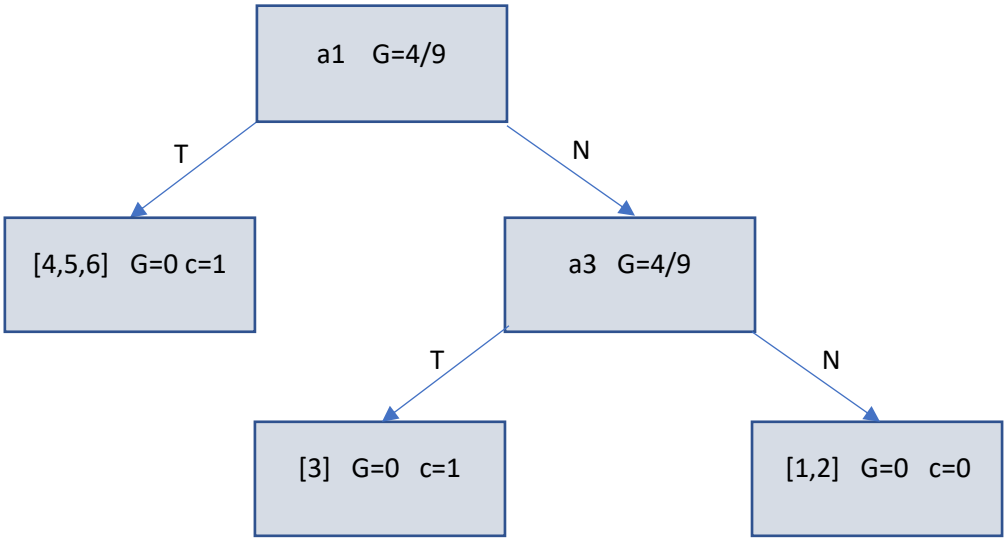
$$G_{2a_3} = 1 - [1^2 + 0^2] = 1 - 1 = 0$$

$$G'_{2a_3} = 1 - [0^2 + 1^2] = 1 - 1 = 0$$

$$IG_{a_3} = \frac{4}{9} - \left(\frac{2}{3} \cdot 0 + \frac{1}{3} \cdot 0\right) = \frac{4}{9} - 0 = \frac{4}{9}$$

$IG_{a_3} > IG_{a_2}$ - wybieramy atrybut a_3

Drzewo decyzyjne po kroku 2):



Opis danych

Dwa niezależne zbiory danych, które w formie pliku .dat znajduję się tutaj i tutaj, opisują kolejno przykłady donorów i akceptorów. Każdą sekwencję DNA poprzedza jej klasa, pozytywna (1) lub negatywna (0), w oparciu o którą przebiegać będzie klasyfikacja. Klasa ta oznacza obecność rozcięcia sekwencji kodującej DNA w danym przykładzie. Dane należące do zbiorów uczącego i weryfikującego zostaną dobrane losowo, lecz ze zwróceniem uwagi na odpowiedni stosunek ilości przykładów pozytywnych i negatywnych w danej grupie. Ich rozmiar będzie jednym z parametrów modyfikacji podlegających ocenie.

Miary jakości

Planowane jest określenie jakości rezultatu przy pomocy:

- Tablicy pomyłek (ang. Confusion matrix)
- Błędu klasyfikacji na zbiorze trenującym
- Dokładności jako procent poprawnych predykcji w stosunku do całkowitej liczby próbek
- Precyzji jako stosunek liczby prawdziwie pozytywnych predykcji do sumy prawdziwie pozytywnych i fałszywie pozytywnych predykcji
- Czułości jako stosunek liczby prawdziwie pozytywnych predykcji do sumy prawdziwie pozytywnych i fałszywie negatywnych predykcji
- F1-score, czyli średniej harmonicznej precyzji i czułości. Pozwoli nam na ocenę w której chcemy zwrócić uwagę na istotę równowagi między precyzją a czułością.

Środowisko implementacji

Projekt tworzony będzie przy użyciu języka Python. W celu implementacji algorytmu lub porównania z już istniejącymi, planowane jest wykorzystanie bibliotek takich jak:

- **Numpy** to biblioteka do obliczeń numerycznych w języku Python. Zapewnia efektywne struktury danych do pracy z dużymi, wielowymiarowymi tablicami i macierzami przez co znacznie ułatwi implementacje własnego algorytmu.
- **Scikit-learn** to biblioteka do uczenia maszynowego w Pythonie, która oferuje narzędzia do klasyfikacji, regresji, klastrowania, redukcji wymiarów i wielu innych zadań związanych z uczeniem maszynowym. Posiada ona implementacje drzew decyzyjnych zarówno dla zadania klasyfikacji jak i regresji, wykorzystamy ją więc do porównania wyników eksperymentów z naszym algorytmem.
- **Graphviz** to narzędzie do wizualizacji grafów. W kontekście uczenia maszynowego, Graphviz jest często używane do wizualizacji struktury drzew decyzyjnych wygenerowanych przez modele, takie jak te w bibliotece scikit-learn.
- **Matplotlib** to biblioteka do tworzenia wykresów i wizualizacji danych w Pythonie. Jest ona często używana w połączeniu z innymi bibliotekami do prezentacji wyników analiz danych, do czego również planujemy ją wykorzystać.

Oczywiście wraz z pracą nad projektem nie wykluczamy wykorzystania innych bibliotek, które okażą się przydatne i interesujące.

Plan eksperymentów

- Implementacja własnego algorytmu oraz badanie jego skuteczności dla:
    - Różnych kryteriów stopu:
        - Osiągnięcie czystego zbioru przykładu
        - Głębokości drzewa decyzyjnego
        - Zbyt mała ilość przykładów w konkretnym węźle
    - Różnych współczynników miar atrybutów pod względem zapewnienia najlepszej skuteczności modelu:
        - Miara Entropii
        - Współczynnik Gini’ego
    - Różnej wielkości zbioru uczącego i weryfikującego
- Porównanie gotowych algorytmów z własną implementacja, wykorzystując dostępne modyfikacje i parametryzacje