# Drzewa decyzyjne.

## 1. Wprowadzenie.

Drzewa decyzyjne są graficzną metodą wspomagania procesu decyzyjnego. Jest to jedna z najczęściej wykorzystywanych technik analizy danych. Drzewo składają się z korzenia oraz gałęzi prowadzących z korzenia do kolejnych wierzchołków. Wierzchołki, z których wychodzi co najmniej jedna krawędź, są nazywane węzłami, a pozostałe wierzchołki – liśćmi. W każdym węźle sprawdzany jest pewien warunek dotyczący danej obserwacji, i na jego podstawie wybierana jest jedną z gałęzi prowadząca do kolejnego wierzchołka. Klasyfikacja danej obserwacji polega na przejściu od korzenia do liścia i przypisaniu do tej obserwacji klasy zapisanej w danym liściu. Poniżej przedstawiono przykładowe drzewo decyzyjne wspomagające zakup konsoli do gier.

## 2. Zastosowanie drzew decyzyjnych

Drzewa decyzyjne znajdują praktyczne zastosowanie w różnego rodzaju problemach decyzyjnych, szczególnie takich gdzie występuje dużo rozgałęziających się wariantów a także w warunkach ryzyka. Wiele algorytmów uczenia się wykorzystuje drzewa decyzyjne do reprezentacji hipotez. Zgodnie z ogólnym celem uczenia się indukcyjnego, dążą one do uzyskania drzewa decyzyjnego klasyfikującego przykłady trenujące z niewielkim błędem próbki i o możliwie niewielkim rozmiarze, w nadziei, że takie drzewo będzie miało również niewielki błąd rzeczywisty. Drzewa decyzyjne znajdują szerokie zastosowanie w problemach związanych z klasyfikacją i predykcją pojęć typu:

- diagnostyka medyczna,
- przewidywanie wydajności,
- akceptacja i udzielanie kredytów.
- i wiele więcej

Proces klasyfikacji z wykorzystaniem drzew decyzyjnych jest efektywny obliczeniowo, wyznaczenie kategorii przykładu wymaga w najgorszym razie przetestowania raz wszystkich jego atrybutów.

## 3. Przykład 1.

Rozważmy sytuację w której decydent podejmuje działania, których wynik zależy od okoliczności od niego niezależnych, ale na których wyniki może reagować podejmując kolejne działania. Powiemy wówczas, że decydent podejmuje decyzje sekwencyjne w warunkach niepewności. Proces podejmowania decyzji sekwencyjnych wygodnie jest przedstawić w postaci drzewa decyzyjnego. W drzewie wyróżniamy :

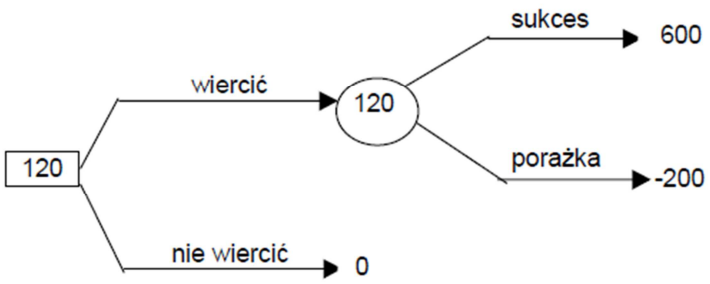
- węzły decyzyjne (kwadraty) reprezentujące decyzje,
- wierzchołki (kółka) reprezentujące zdarzenia losowe.

Łuki wychodzące z węzłów decyzyjnych będziemy utożsamiać z podjętymi decyzjami, a łuki wychodzące z wierzchołków odpowiadających zdarzeniom losowym z wynikami jakie wystąpią w przypadku zajścia zdarzeń losowych wpływających na proces decyzyjny (wynik podjętej decyzji). Wewnątrz wierzchołków – wypłaty, które uzyskujemy w kolejnych etapach procesu decyzyjnego.
Poszukiwacz ropy musi podjąć decyzję, czy rozpocząć wiercenie szybu naftowego w pewnym miejscu przed wygaśnięciem licencji na wykonywanie odwiertów. Koszt wiercenia wynosi 200 tys. dol., całkowite zyski (bez uwzględnienia kosztów wierceń) w przypadku natrafienia na ropę – 800 tys. dol.

Decyzja o podjęciu wierceń jest pozytywna, jeżeli oczekiwany zysk związany z podjęciem decyzji jest dodatni.

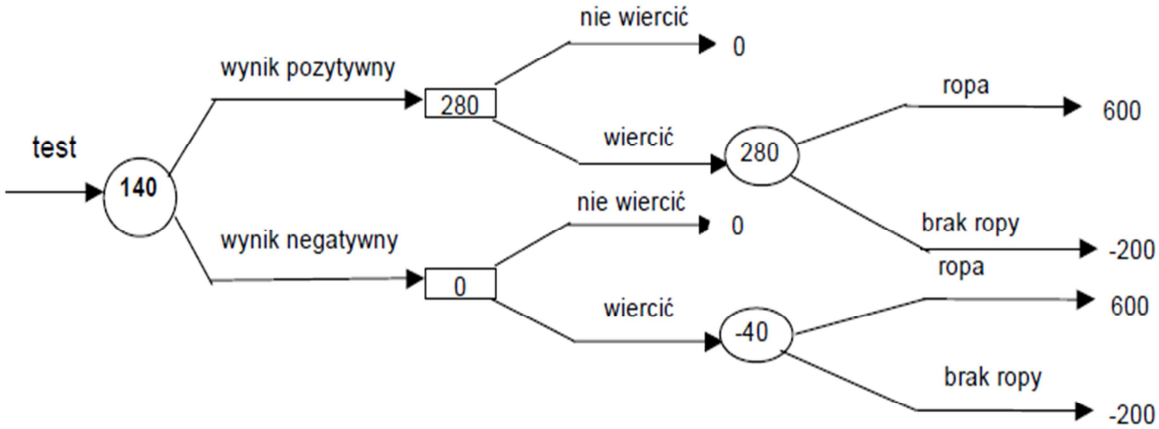
Jeżeli decyzja jest podejmowana wielokrotnie, to wybór kryterium maksymalizacji oczekiwanego zysku jest w pełni uzasadniony.

Drzewo decyzyjne problemu wyboru strategii wierceń




Powiedzmy, że przed dokonaniem odwiertu możemy przeprowadzić test sejsmiczny pozwalający na bardziej precyzyjną ocenę warunków geologicznych działki. Test ten pozwala z większą dokładnością odpowiedzieć na pytanie, czy trafimy na ropę. Oznacza to zwiększenie prawdopodobieństwa trafienia i zmniejszenie prawdopodobieństwa nietrafienia na ropę – a więc zwiększenie wartości oczekiwanej wypłaty (zysku). Przeprowadzenie testu wiąże się z poniesieniem dodatkowych kosztów. Kiedy warto je ponieść?

Drzewo decyzyjne problemu pozyskania dodatkowej informacji




## 4. Przykład 2.

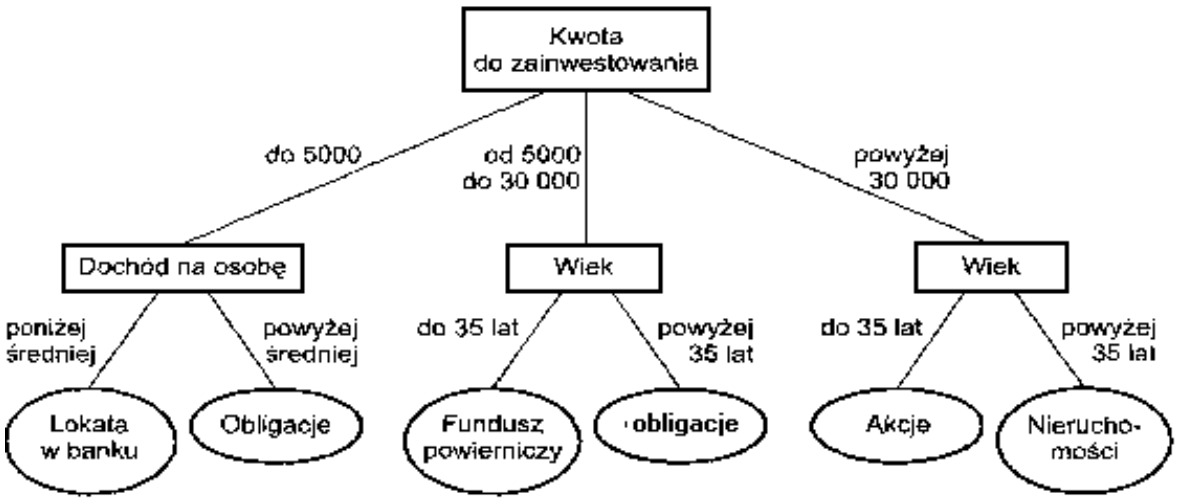
Rozważmy prosty system ekspertowy ze ścisłe określonym zbiorem rozwiązań i kilkoma możliwościami wyboru. Nasz system ekspertowy będzie doradzał, gdzie ulokować pewną kwotę oszczędności. Decyzja o sposobie zainwestowania zależy od wysokości kwoty, dochodów na osobę i wieku inwestora.

Jedną z metod pozwalających łatwo i szybko stworzyć bazę reguł jest zaprojektowanie drzewa decyzyjnego dla rozwiązywanego problemu. Każdy węzeł takiego drzewa reprezentuje albo pytanie o wartość atrybutu, albo konkluzję. Każda gałąź wywodząca się z poszczególnego węzła związanego z pytaniem o atrybut reprezentuje jedną z możliwych

wartości tego atrybutu. Na rysunku przedstawiającym drzewo decyzyjne naszego problemu węzły związane z pytaniem będą oznaczone prostokątami, a węzły dotyczące konkluzji - owalami. Załóżmy że mamy następującą listę atrybutów i ich wartości:

1. *Inwestycje:* lokata w banku, obligacje, fundusz powierniczy, akcje, nieruchomości.
2. *Kwota do zainwestowania:* do 5000, od 5000 do 30000, powyżej 30000.
3. *Dochód na osobę:* poniżej średniej krajowej, powyżej średniej krajowej.
4. *Wiek inwestora:* do 35 lat, powyżej 35 lat.

Drzewo decyzyjne dla tego przykładu prezentuje rysunek poniżej. Przyjmujemy dla tego problemu węzeł *Kwota do zainwestowania* jako węzeł główny, czyli leżący na najwyższym poziomie drzewa, zawierający pod sobą wszystkie możliwe rozwiązania. Na następnym poziomie drzewa wprowadzimy atrybuty *Dochód na osobę* oraz *Wiek.* Możemy teraz przystąpić do tworzenia reguł na podstawie utworzonego drzewa. Przykładowe drzewo nie zawiera hipotez pośrednich, co uprości proces tworzenia reguł. Tworząc regułę, wybieramy węzeł konkluzji, który do tej pory nie był rozpatrywany. Śledzimy ścieżkę w górę drzewa decyzyjnego przez wszystkie węzły aż do węzła głównego. Węzły oznaczone owalami zapisujemy w części THEN (TO) reguły, a oznaczone prostokątami w części IF (JEŻELI). Każda ścieżka utworzy nam jedną regułę w bazie wiedzy. Na podstawie naszego drzewa decyzyjnego powstaną następujące reguły o nazwach oznaczonych kolejnymi cyframi:



Reguła l: **IF** *kwota do zainwestowania* = do 5000

**and** *dochód na osobę* = poniżej średniej

**THEN** *inwestycja* = lokata w banku.

Reguła 2: **IF** *kwota do zainwestowania* = do 5000

**and** *dochód na osobę* = powyżej średniej

**THEN** *inwestycja* = obligacje.

Reguła 3: **IF** *kwota do zainwestowania* = od 5000 do 30000

**and** *wiek* = do 35 lat

**THEN** *inwestycja* = fundusz powierniczy.

Reguła 4: **IF** *kwota do zainwestowania* = od 5000 do 30000

**and** *wiek* = powyżej 35 lat

**THEN** *inwestycja* = obligacje.

Reguła 5: **IF** *kwota do zainwestowania* = powyżej 30000

**and** *wiek* = do 35 lat

**THEN** *inwestycja* = akcje.

Reguła 6: **IF** *kwota do zainwestowania* = powyżej 30000

**and** *wiek* = powyżej 35 lat

**THEN** *inwestycja* = nieruchomości.

Dla zdefiniowanego przez nas problemu otrzymaliśmy kompletną i spójną bazę wiedzy.

Powracając do naszego przykładu, załóżmy, że przy doradzaniu o sposobie inwestowania oszczędności chcemy uzyskać dokładniejsze informacje o inwestorze. Dołożymy następujące atrybuty i ich wartości:

Dla inwestorów pragnących zainwestować kwotę do 5000:

• *Przeinaczenie kwoty:* niewielki zakup, większy zakup (np. samochód lub zmiana mieszkania.

• *Termin wykorzystania:* do roku, minimum za rok. Dla pozostałych inwestorów:

• *Ryzyko:* tak, nie.

• *Stan majątkowy:* dobry, zły.

Nasze drzewo ulegnie teraz modyfikacji i będzie miało postać jak na rysunku 2. Zauważmy, że hipotezy w naszym drzewie pojawiają się teraz na różnych poziomach drzewa.

Tworząc jak poprzednio zbiór reguł na podstawie siedzenia poszczególnych gałęzi, możemy otrzymać zbiór reguł składających się na bazę wiedzy. Dla inwestorów inwestujących do 5000 zbiór reguł ma postać następującą:

**Reguła l:** **IF** *kwota do zainwestowania* = do 5000

**and** *dochód na osobę* = poniżej średniej

**THEN** *inwestycja* = lokata w banku.

**Reguła 2:** **IF** *kwota do zainwestowania* = do 5000

**and** *dochód na osobę* = powyżej średniej

**and** *przeznaczenie* = niewielki zakup

**THEN :** *inwestycja* = lokata w banku.

**Reguła 3:** **IF** *kwota do zainwestowania* = do 5000

**and** *dochód na osobę* = powyżej średniej

**and** *przeznaczenie* = większy zakup

**and** *termin wykorzystania* = do roku

**THEN** *inwestycja* = fundusz powierniczy.

**Reguła 4:** **IF** *kwota do zainwestowania* = do 5000

**and** *dochód na osobę* = powyżej średniej

**and** *przeznaczenie* = większy zakup

**and** *termin wykorzystania* = minimum za rok

**THEN** *inwestycja* = obligacje.

Analogicznie możemy zbudować reguły dla pozostałych konkluzji. W przypadku, gdy reguły są długie lub gdy nie jesteśmy w stanie zbudować drzewa decyzyjnego, możemy posłużyć się hipotezami pośrednimi, będącymi prostym sposobem określenia części procesu decyzyjnego. W naszym przykładzie hipotezą pośrednią może być atrybut *Typ klienta. Typ klienta* określimy na podstawie *Kwoty* do zainwestowania i *Wieku* (patrz tablica 1).

**Tablica** 1. Określenie typu klienta

| Typ klienta | Wiek | Kwota |
|---|---|---|
| A | do 35 lat | 5000-30000 |
| B | powyżej 35 | 5000-30000 |
| C | do 35 lat | powyżej 30000 |
| D | powyżej 35 | powyżej 30000 |

Reguły dotyczące inwestycji powyżej 5000 będziemy tworzyć za pomocą hipotezy pośredniej. Tak więc reguły bazy definiujące podcel mają następującą postać:

**Reguła 5:**

**IF** *kwota do zainwestowania* = od 5000 do 30000

**and** *wiek = do* 35 lat

**THEN** *typ klienta* = A.

**Reguła 6:**

**IF** *kwota do zainwestowania* = od 5000 do 30000

**and** *wiek* = powyżej 35 lat

**THEN** *typ klienta* = B.

**Reguła 7:**

**IF** *kwota do zainwestowania* = powyżej 30000

**And** *wiek = do* 35 lat

**THEN** *typ klienta = C.*

**Reguła 8:**

**IF** *kwota do zainwestowania* = powyżej 30000

**and** *wiek* = powyżej 35 lat

**THEN** *typ klienta = D.*

Korzystając z zdefiniowanych hipotez pośrednich, tworzymy pozostałe reguły potrzebne nam do wypełnienia bazy danych. W regułach tych konkluzje pośrednie reguł 5, 6, 7 i 8 są przesłankami:

**Reguła 9:**

IF ***typ klienta =*** **A**

and ***ryzyko =*** **tak**

THEN ***inwestycja =*** **akcje.**

**Reguła 10:**

IF ***typ klienta*** **- A**

and ***ryzyko =*** **nie**

THEN ***inwestycja =o*****bligacje.**

**Reguła 11:**

IF ***typ klienta =*** **B**

and ***ryzyko =*** **tak**

THEN ***inwestycja =*** **fundusz powierniczy.**

**Reguła 12:**

IF ***typ klienta =*** **B**

and ***ryzyko*** **= nie**

THEN ***inwestycja =*** **lokata w banku.**

**Reguła 13:**

IF ***typ klienta = C or typ klienta =*** **D**

and ***stan majątkowy =*** **dobry**

and ***ryzyko =*** **tak**

THEN ***inwestycja =*** **akcje.**

**Reguła 14:**

IF ***typ klienta =*** **C or** ***typ klienta =*** **D**

and ***stan majątkowy =*** **zły**

THEN ***inwestycja =*** **fundusz powierniczy.**

**Reguła 15:**

IF ***typ klienta = C***

and ***stan majątkowy*** **= dobry**

and ***ryzyko*** **— nie**

THEN ***inwestycja =*** **fundusz powierniczy.**

**Reguła 16:**

IF ***typ klienta =*** **D**

and ***stan majątkowy =*** **dobry**

and ***ryzyko*** **= nie**

THEN ***inwestycja =*** **nieruchomości.**

Przy tworzeniu drzewa decyzyjnego może okazać się, iż posiada ono gałęzie prowadzące do nieistniejących lub niemożliwych do określenia rozwiązań. Należy eliminować takie sytuacje poprzez reorganizowanie węzłów drzewa. Lepiej tworzyć drzewa decyzyjne z jak najmniejszą liczbą poziomów (tym samym atrybutów), reprezentujące wszystkie znane konkluzje, a tym samym tworzące bardziej efektywne reguły. Im mniejsza jest liczba atrybutów, tym mniejsza liczba danych potrzebna do uzyskania konkluzji i mniejsza liczba pytań zadawanych użytkownikowi. Zwłaszcza w przypadkach deterministycznych. Należy jednak uważać, by w trakcie modyfikacji nie powstawały sytuacje, w których wymagany będzie dodatkowy atrybut, wcześniej nie uwzględniony. W tych przypadkach, w których mamy do czynienia z niepewnością, lepiej jest posiadać dokładniejsze informacje przy podejmowaniu decyzji.

Prześledźmy teraz sposób wnioskowania w naszym przykładzie. Załóżmy, że znane są następujące fakty:

1. *Kwota do zainwestowania:* 35000 (powyżej 30000).

2. *Wiek:* 33 łata (do 35 lat).

3. S*tan majątkowy:* dobry.

4. *Ryzyko',* podejmuje niechętnie (nie).

Inwestor chce uzyskać poradę, jak zainwestować pieniądze. W procesie wnioskowania w przód na podstawie znanych faktów i reguł generujemy nowe fakty aż do osiągnięcia celu lub niemożności uzyskania rozwiązania z powodu braku reguł. Rozpoczynamy od rozważenia warunków reguł. Szukamy reguł, w których w części IF prawdziwa jest przesłanka *Kwota do zainwestowania =* powyżej 30000. W naszym przypadku są to reguły 7 i 8.

Reguła 7 jako konkluzję zawiera *Typ klienta* = C. Musimy zweryfikować tę konkluzję, sprawdzając prawdziwość przesłanki *Wiek* = do 35 lat. Ta przesłanka jest prawdziwa, więc konkluzja *Typ klienta = C* też jest prawdziwa. W przypadku reguły 8 weryfikacja konkluzji *Typ klienta* = D jest nieprawdziwa, gdyż przesłanka *Wiek* = powyżej 35 lat jest nieprawdziwa. Wiedząc, że *Typ klienta* = C, możemy zweryfikować reguły, w których występuje on jako przesłanka. Są to reguły 13, 14, 15.

W regule 13 prawdziwe są przesłanki *Typ klienta* = C i *Stan majątkowy =* dobry. Przesłanka *Ryzyko =* tak jest nieprawdziwa, a tym samym akcja reguły jest nieprawdziwa. W regule 14 przesłanka *Typ klienta = C* jest prawdziwa, natomiast przesłanka *Stan majątkowy* = zły, jest fałszywa, co powoduje fałszywość konkluzji. W regule 15 przesłanka *Typ klienta* = C jest prawdziwa, przesłanka *Stan majątkowym* dobry jest prawdziwa oraz przesłanka *Ryzyko =* nie także jest prawdziwa, a tym samym konkluzja reguły jest prawdziwa. Konkluzja tej reguły *Inwestycja =* fundusz powierniczy jest konkluzją ostateczną i tak też brzmi porada dla inwestora.

Na tym samym przykładzie prześledźmy teraz sposób wnioskowania w tył. Załóżmy, że mamy następujące fakty:

1. *Kwota do zainwestowania:* 35000 (powyżej 30000).

2. *Wiek:* 53 lata (powyżej 35 lat).

3. *Stan majątkowy:* dobry.

4. *Ryzyko:* podejmuje niechętnie (nie).

Inwestor chce się dowiedzieć, czy powinien inwestować w nieruchomości. Wnioskowanie do tyłu odbywa się przez akcje reguł. Rozpoczynamy od reguły, której akcja rozwiązuje problem. W naszym przypadku hipoteza *Inwestycja* == nieruchomości występuje w regule 16. Musimy sprawdzić jej prawdziwość, weryfikując wszystkie przesłanki. Zaczynamy od przesłanki *Typ klienta =* D. Ponieważ nie jest to znany nam fakt szukamy reguł, dla których *Typ klienta* = D jest konkluzją reguły. Sytuacja taka występuje w regule 8.

Sprawdzamy teraz wartości przesłanek reguły 8:

• przesłanka *Kwota do zainwestowania =* powyżej 30000 jest prawdziwa,

• przesłanka *Wiek =* powyżej 35 lat jest prawdziwa. Zatem konkluzja *Typ klienta* = D jest prawdziwa.

Powracamy do reguły 16 i weryfikowania kolejnych jej warunków:

• przesłanka *Stan majątkowy =* dobry jest prawdziwa,

• przesłanka *Ryzyko* = nie jest prawdziwa,

wszystkie warunki reguł 16 są więc prawdziwe, co powoduje prawdziwość hipotezy

*Inwestycje* = nieruchomości.

## 5. Podsumowanie

**Zalety**

· Drzewa decyzyjne mogą reprezentować dowolnie złożone pojęcia pojedyncze lub wielokrotne, jeżeli tylko ich definicje da się wyrazić w zależności od atrybutów. · Efektywność pamięciowa reprezentacji drzewiastej. · Czas decyzyjny ograniczony liniowo przez liczbę atrybutów (maksymalna głębokość drzewa).

· Forma reprezentacji czytelna dla człowieka.

· Łatwość przejęcia od reprezentacji drzewiastej do reprezentacji regułowej.

**Wady**

· Testuje się wartość jednego atrybutu na raz, co powoduje niepotrzebny rozrost drzewa dla danych gdzie poszczególne atrybuty zależą od siebie (inne metody reprezentacji mogą być w tym przypadku o wiele mniej złożone).

· Kosztowna reprezentacja alternatyw pomiędzy atrybutami – znaczny rozrost drzewa (w przeciwieństwie do reprezentacji koniunkcji, która jest zapisywana jako pojedyncza „ścieżka”, czyli droga od korzenia do liścia).

· Drzewa decyzyjne nie stwarzają łatwej możliwości do ich inkrementacyjnego aktualizowania, algorytmy udoskonalające gotowe już drzewa poprzez zestaw nowych przykładów są bardzo złożone i zazwyczaj wynikiem jest drzewo gorszej jakości niż drzewo budowane od początku z kompletnym zestawem przykładów.