Uczenie Maszynowe
Dokumentacja wstępna

Michał Matak, Jakub Robaczewski

Temat:

Tworzenie drzewa decyzyjnego przy pomocy algorytmu ewolucyjnego. Zwykle klasyfikatory budowane są w oparciu o metodę zachłanną - w kolejnym kroku wybieramy lokalnie najlepszy podział. Takie podejście jest bardzo szybkie jednak nie zawsze prowadzi do utworzenia optymalnej struktury drzewa.

Krótki opis projektu:

W projekcie zostanie zbadana możliwość tworzenia drzewa decyzyjnego za pomocą algorytmu ewolucyjnego i porównanie efektów z drzewami utworzonymi na podstawie metody zachłannej. W celu porównania zostaną przeprowadzane eksperymenty na 6 zbiorach danych. Algorytmy, eksperymenty oraz zbiory danych zostały opisane poniżej.

Opis algorytmów:

Algorytm zachłanny (punkt wyjścia):

Konstrukcja drzewa za pomocą klasycznego algorytmu zachłannego wygląda w następujący sposób.

Mając zbiór danych w węźle wybieramy podział, według którego nastąpi rozgałęzienie drzewa. Podziałem może być np.:

- dla atrybutów dyskretnych:
  - podział binarny – przyporządkowanie do jednego podzbioru jeśli atrybut jest równy jakiejś wartości, do drugiego jeśli nie jest równy
  - podział wielowartościowy na podstawie podzbiorów wartości atrybutu – przyporządkowanie do jakiegoś podzbioru z *n* podzbiorów jeśli atrybut jest równy danej dla niego wartości
  - podział binarny na podstawie przynależności do zbioru - przyporządkowanie do jednego podzbioru jeśli atrybut jest równy należy do jakiegoś podzbioru wartości, do drugiego jeśli nie należy
- dla atrybutów ciągłych
  - podział binarny na podstawie nierówności - przyporządkowanie do jednego podzbioru jeśli atrybut jest mniejszy lub równy od jakiejś wartości, do drugiego w przeciwnym przypadku
  - podział wielowartościowy na podstawie przedziałów wartości atrybutu - przyporządkowanie do jakiegoś podzbioru z *n* podzbiorów jeśli atrybut należy do danego dla niego przedziału

Z możliwych zdefiniowanych podziałów wybierany jest ten podział, dla którego wartość nieczystości jest najmniejsza, czyli taki dla którego wartość entropii warunkowej jest najmniejsza. Wartość entropii warunkowej wylicza się równaniem:

$$E_{T_n}(c|t) = \sum_{r \epsilon R_t} \frac{|T_{n,t=r}|}{|T_n|} E_{T_{n,t=r}}(c)$$

Gdzie:

- t to rozważany podział
- n to obecnie rozpatrywany węzeł
- r to wynik podziału t
- T_n to zbiór przykładów trenujących w danym węźle
- $T_{n,t=r} = \{x \in T_n | \, t(x) = r\}$, czyli podzbiór przykładów trenujących powstały przy podziale t dla danego wyniku podziału r
- $|T_n|$ to liczba elementów zbioru T_n, (analogicznie $|T_{n,t=r}|$ to liczba elementów $T_{n,t=r}$)

- *c* to *pojęcie*, czyli skończony zbiór kategorii - jest to dość abstrakcyjne określenie, jednak należy tylko wiedzieć, iż jego entropię można policzyć za pomocą:

$$-\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

gdzie $p(x_i)$ jest prawdopodobieństwem zajścia zdarzenia x_i (inaczej prawdopodobieństwo znalezienia w zbiorze wartości x_i, gdzie x_i to jedna z wartości, które model może zwracać na wyjściu)

W praktyce tworzenie drzewa zaczyna się od pełnego zbioru trenującego w korzeniu, dla którego wybiera się podział minimalizujący wartość entropii warunkowej. Następnie na podstawie tego podziału należy stworzyć węzły będące dziećmi obecnie rozpatrywanego węzła (w tym przypadku korzenia) i przydzielić do nich podzbiory obecnego zbioru w węźle według wybranego podziału. Operacje te należy powtarzać rekurencyjnie w węzłach-dzieciach aż do momentu, gdy w węzłach utworzonych przez podział przykłady treningowe będą należały do tylko jednej klasy (te węzły stają się wtedy liśćmi).

Wizualizacja krok po kroku

Aby zaprezentować tworzenie drzewa decyzyjnego został użyty zbiór danych pokazany na wykładzie. Do niego dodany został kompletny opis tworzenia całego drzewa. Atrybuty są dyskretne i mogą przyjmować kilka wartości dlatego zostanie zastosowany podział wielowartościowy.

x	*outlook*	*temperature*	*humidity*	*wind*	*play*
1	*sunny*	*hot*	*high*	*normal*	*no*
2	*sunny*	*hot*	*high*	*high*	*no*
3	*overcast*	*hot*	*high*	*normal*	*yes*
4	*rainy*	*mild*	*high*	*normal*	*yes*
5	*rainy*	*cold*	*normal*	*normal*	*yes*
6	*rainy*	*cold*	*normal*	*high*	*no*
7	*overcast*	*cold*	*normal*	*high*	*yes*
8	*sunny*	*mild*	*high*	*normal*	*no*
9	*sunny*	*cold*	*normal*	*normal*	*yes*
10	*rainy*	*mild*	*normal*	*normal*	*yes*
11	*sunny*	*mild*	*normal*	*high*	*yes*
12	*overcast*	*mild*	*high*	*high*	*yes*
13	*overcast*	*hot*	*normal*	*normal*	*yes*
14	*rainy*	*mild*	*high*	*high*	*no*

Mając powyższą tabelkę rozważany jest podział ze względu na jeden z atrybutów. Dla każdego podziału jest liczona entropia warunkowa, na podstawie której zostanie znaleziony optymalny podział w korzeniu:

- *outlook*

$$E_{T_{outlook=sunny}}(c) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \approx 0.9710$$

$$E_{T_{outlook=overcast}}(c) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

$$E_{T_{outlook=rainy}}(c) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \approx 0.9710$$

$$E_T(c|outlook) = \frac{5}{14}\cdot 0.9710 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0.9710 \approx 0.694$$

- *temperature*

$$E_{T_{temperature=hot}}(c) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$E_{T_{temperature=mild}}(c) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} \approx 0.9183$$

$$E_{T_{temperature=cold}}(c) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} \approx 0.8113$$

$$E_T(c|temperature) = \frac{4}{14}\cdot 1 + \frac{6}{14}\cdot 0.9183 + \frac{4}{14}\cdot 0.8113 \approx 0.9111$$

- *humidity*

$$E_{T_{humidity=high}}(c) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} \approx 0.9852$$

$$E_{T_{humidity=normal}}(c) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} \approx 0.5917$$

$$E_T(c|humidity) = \frac{7}{14}\cdot 0.9852 + \frac{7}{14}\cdot 0.5917 \approx 0.7885$$

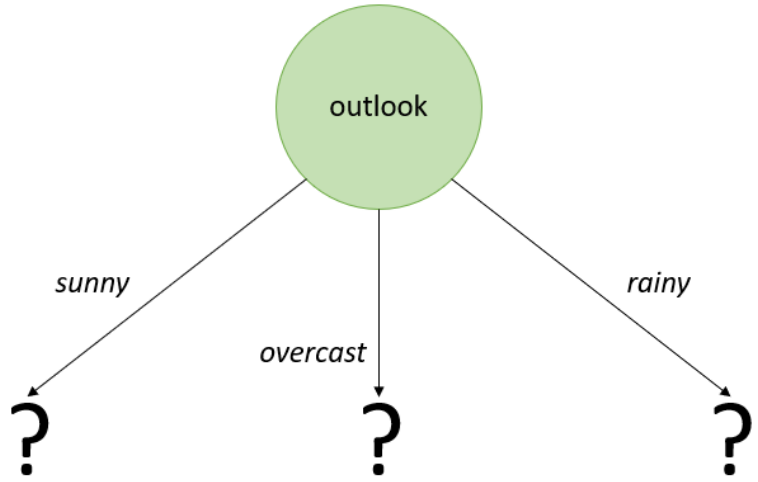
- *wind*

$$E_{T_{wind=high}}(c) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

$$E_{T_{wind=normal}}(c) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} \approx 0.8113$$

$$E_T(c|wind) = \frac{6}{14}\cdot 1 + \frac{8}{14}\cdot 0.8113 \approx 0.8922$$

Wartość entropii warunkowej jest najmniejsza dla podziału $E_T(c|outlook)$, zatem ten podział zostanie dokonany według atrybutu *outlook* i zostaną utworzone trzy węzły-dzieci.



Rozpatrując teraz sytuacje w trzech węzłach:

- Węzeł *outlook = sunny*

x	*outlook*	*temperature*	*humidity*	*wind*	*play*
1	*sunny*	*hot*	*high*	*normal*	*no*
2	*sunny*	*hot*	*high*	*high*	*no*
8	*sunny*	*mild*	*high*	*normal*	*no*
9	*sunny*	*cold*	*normal*	*normal*	*yes*
11	*sunny*	*mild*	*normal*	*high*	*yes*

Ponownie obliczanie entropii warunkowej dla atrybutów:

  - *temperature*

$$E_{T_{temperature=hot}}(c) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$$

$$E_{T_{temperature=mild}}(c) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$E_{T_{temperature=cold}}(c) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$$

$$E_T(c|temperature) = \frac{2}{5}\cdot 0 + \frac{2}{5}\cdot 1 + \frac{1}{5}\cdot 0 = 0.4$$

- *humidity*

$$E_{T_{humidity=high}}(c) = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = 0$$

$$E_{T_{humidity=normal}}(c) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$$

$$E_T(c|humidity) = \frac{3}{5}\cdot 0 + \frac{2}{5}\cdot 0 = 0$$

- *wind*

$$E_{T_{wind=high}}(c) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$E_{T_{wind=normal}}(c) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} \approx 0.9183$$

$$E_T(c|wind) = \frac{2}{5}\cdot 1 + \frac{3}{5}\cdot 0.9183 \approx 0.9501$$

W tym przypadku wartość entropii warunkowej jest najmniejsza dla $E_T(c|humidity)$, zatem podział zostanie dokonany według parametru *humidity*. Można także od razu stwierdzić, że *play* będzie przyjmuje wartość *no* dla *humidity* równego *high*, natomiast dla *normal* będzie przyjmować wartość *yes*. Powstałe węzły-dzieci staną się liśćmi.

- Węzeł *outlook = overcast*

x	*outlook*	*temperature*	*humidity*	*wind*	*play*
3	*overcast*	*hot*	*high*	*normal*	*yes*
7	*overcast*	*cold*	*normal*	*high*	*yes*
12	*overcast*	*mild*	*high*	*high*	*yes*
13	*overcast*	*hot*	*normal*	*normal*	*yes*

W tym przypadku wartość *play* zawsze wynosi *yes*. Zatem ten węzeł można uczynić liściem zwracającym *yes*.

- Węzeł *outlook = rainy*

x	*outlook*	*temperature*	*humidity*	*wind*	*play*
4	*rainy*	*mild*	*high*	*normal*	*yes*
5	*rainy*	*cold*	*normal*	*normal*	*yes*
6	*rainy*	*cold*	*normal*	*high*	*no*
10	*rainy*	*mild*	*normal*	*normal*	*yes*
14	*rainy*	*mild*	*high*	*high*	*no*

Ponownie obliczanie entropii warunkowej dla atrybutów:

- *temperature*

$$E_{T_{temperature=mild}}(c) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \approx 0.9183$$

$$E_{T_{temperature=cold}}(c) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$E_T(c|temperature) = \frac{3}{5}\cdot 0.9183 + \frac{2}{5}\cdot 1 = 0.9501$$

- *humidity*

$$E_{T_{humidity=high}}(c) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$E_{T_{humidity=normal}}(c) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \approx 0.9183$$

$$E_T(c|humidity) = \frac{2}{5}\cdot 0 + \frac{3}{5}\cdot 0.9183 = 0.9501$$

- *wind*

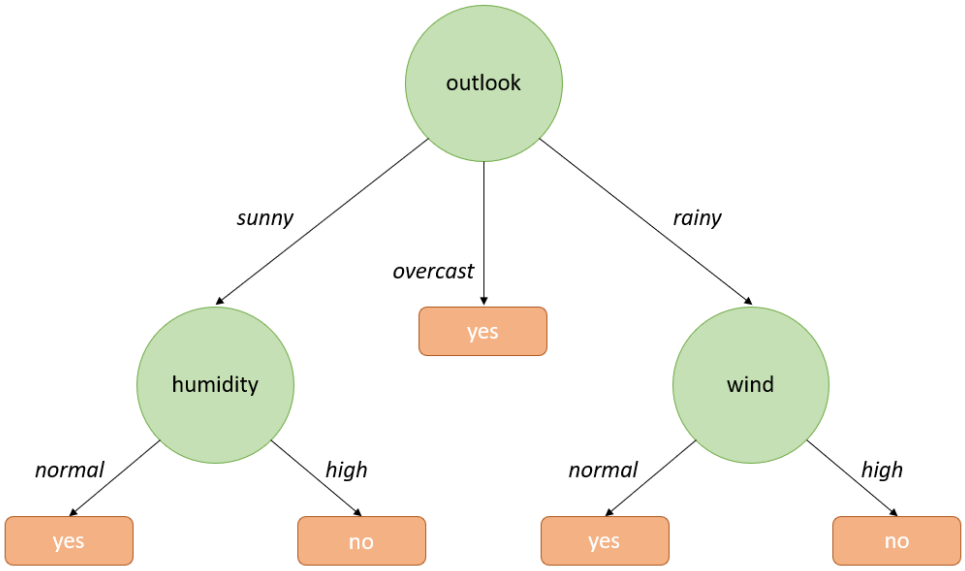
$$E_{T_{wind=high}}(c) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$$

$$E_{T_{wind=normal}}(c) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0$$

$$E_T(c|wind) = \frac{2}{5}\cdot 0 + \frac{3}{5}\cdot 0 = 0$$

W tym przypadku wartość entropii warunkowej jest najmniejsza dla $E_T(c|wind)$, zatem podział zostanie dokonany według parametru *wind*. Można także od razu stwierdzić, że *play* będzie przyjmować wartość *no* dla *wind* równego *high*, natomiast dla *normal* będzie przyjmować wartość *yes*. Powstałe węzły-dzieci staną się liśćmi.

Ostatecznie drzewo decyzyjnie będzie miało postać:



Algorytm ewolucyjny:

Przy tworzeniu drzewa decyzyjnego za pomocą algorytmu ewolucyjnego sama struktura algorytmu jest taka sama jak w przypadku każdego algorytmu ewolucyjnego:

1. P_0 <- Inicjalizacja populacji początkowej
2. *najlepszyOsobnik* <- wybranie najlepszego spośród (P_t i N)
3. $t = 0$
4. Dopóki nie zostało spełnione kryterium stopu:
   I. O <- reprodukcja populacji (P_t)
   II. C <- krzyżowanie populacji(O)
   III. M <- mutacja populacji(C)
   IV. N <- ewaluacja populacji(M)
   V. *najlepszyOsobnik* <- wybranie najlepszego spośród (P_t i N)
   VI. P_{t+1} <- sukcesja(P_t i N)
   VII. $t = t + 1$

W tym podejściu drzewo jest reprezentowane przez jednego osobnika populacji. Aby w pełni zdefiniować algorytm ewolucyjny należy zdefiniować dla osobnika odpowiednie operacje.

Inicjalizacja

Podczas tworzenia osobnika w każdym węźle z prawdopodobieństwem *p* dochodzi do podziału drzewa, w przeciwnym razie węzeł zamienia się w liść. Jeśli doszło do podziału to następuje losowanie atrybutu wedle którego następuje podział (i wedle jakich parametrów). Jeśli węzeł staje się liściem to następuje losowanie na którą klasę wskazuje liść.

Ocena osobnika - funkcja kosztu

Ocena osobnika będzie przebiegać na podstawie wzoru:

$$\alpha \bullet \text{dokładność} + \beta \bullet \text{rozmiar drzewa}$$

Gdzie α oraz β są parametrami wybieranymi przez użytkownika aby móc uwzględnić przy ocenie rozmiar drzewa i na jego podstawie zmniejszać wartość nadmiernie rozbudowanych drzew (pod warunkiem, że parametr β będzie ujemny).

Warunek stopu

Warunkiem stopu będzie przekroczenie określonej liczby iteracji lub brak postępów od pewnej ilości iteracji (np. brak wzrostu oceny lub wzrost marginalny od 10 iteracji).

Reprodukcja

Reprodukcją (inaczej zwana selekcją) jest dobranie osobników z danej populacji do populacji potomnej i przebiega na podstawie oceny osobników. W tym celu zostanie wykorzystana selekcja turniejowa. Wybiera się do niej *k* osobników z populacji rodzicielskiej, a następnie najlepszego z tych osobników umieszcza się w populacji potomnej.

Krzyżowanie

Do krzyżowania potrzebne są dwa osobniki. Podczas krzyżowania z obu osobników wybierany jest losowy węzeł, a następnie poddrzewo (drzewo od danego węzła w dół) z węzła pierwszego osobnika zostaje umieszczone w miejscu węzła drugiego osobnika.

Mutacja

Podczas mutacji wybierany jest losowy węzeł, a następnie zmieniony zostaje w nim warunek podziału na inny losowy.

Sukcesja

Sukcesja decyduje które osobniki przeżyją i trafią do następnego pokolenia. Zostanie zastosowania sukcesja elitarna. Podczas sukcesji elitarnej wybiera się *k* najlepszych osobników, oraz wszystkie osobniki z populacji potomnej z wyłączeniem *k* najgorszych.

Plan eksperymentów:

Na początku zostanie wprowadzona początkowa konfiguracja algorytmu ewolucyjnego, którą będzie

- rozmiar populacji: 20
- $p = 0.3$
- selekcja turniejowa o rozmiarze 5

- α = 100
- β = -1
- sukcesja elitarna z elitą o rozmiarze 1
- kryterium stopu: liczba iteracji >= 500

Następnie nastąpi poszukiwanie parametrów pozwalających na uzyskanie lepszego wyniku.

Głównym celem jednak jest porównanie ze sobą optymalnego algorytmu ewolucyjnego z algorytmem zachłannym. W tym celu dla każdego z niżej wymienionych zbiorów danych za pomocą obu algorytmów zostaną utworzone drzewa decyzyjne, a następnie policzona zostanie ich dokładność. W każdym przypadku algorytm będzie uruchomiany 30 razy, co pozwoli na zebranie zagregowanych wyników (średniej, odchylenia standardowego, maksimum i minimum). Na ich podstawie zostanie przeprowadzone porównanie ze sobą algorytmów.

Zbiory danych:

Wybrano 6 zbiorów danych, na podstawie których będą przebiegać eksperymenty. Opisane są one poniżej wraz z wymienionymi klasami oraz atrybutami. W niektórych przypadkach klasy nie są zbilansowane. Podczas wyboru kierowano się głównie różnorodnością tematyki. Wybrano trzy zbiory, które są dość klasyczne w przypadku zadania klasyfikacji (jakość wina, śmierć na Titanicu, gatunek wina), jeden w którym wszystkie atrybuty mają charakter dyskretny (jakość samochodów) oraz dwa, które mają dość praktyczne zastosowanie (tkanka piersiowa oraz przesyłki).

Zbiór do klasyfikacji tkanki piersiowej

https://www.kaggle.com/datasets/ukveteran/breast-tissue-data-set

6 klas (106 przypadków)

1. Carcinoma (21)
2. Fibro-adenoma (15)
3. Mastopathy (18)
4. Glandular (16)
5. Connective (14)
6. Adipose (22)

9 atrybutów, wszystkie mają charakter ciągły

1. Impedivity (ohm) at zero frequency
2. phase angle at 500 KHz
3. high-frequency slope of phase angle
4. impedance distance between spectral ends
5. area under spectrum
6. area normalized by DA
7. maximum of the spectrum
8. distance between I0 and real part of the maximum frequency point
9. length of the spectral curve

Zbiór do klasyfikacji jakości wina czerwonego na podstawie składu.

https://www.kaggle.com/datasets/sh6147782/winequalityred

6 klas jakości (1596 przypadków)

- 3 (18)
- 4 (52)
- 5 (680)
- 6 (637)
- 7 (199)
- 8 (18)

11 atrybutów, wszystkie mają charakter ciągły

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol

Zbiór do klasyfikacji jakości samochodów

https://www.kaggle.com/datasets/elikplim/car-evaluation-data-set

4 klasy (1728)

- unacc (1210)
- acc (384)
- good (69)
- v-good (65)

6 atrybutów, wszystkie mają charakter dyskretny

1. koszt zakupu [low, med, high, vhigh]
2. koszt utrzymania [low, med, high, vhigh]
3. liczba drzwi [2, 3, 4, 5more]
4. liczba pasażerów [2, 4, more]
5. wielkość bagażnika [small, med, big]
6. klasa bezpieczeństwa [small, med, big]

Binarny zbiór do przewidywania śmierci pasażera na podstawie danych z katastrofy Titanica.
https://www.kaggle.com/competitions/titanic/

2 klasy (418)

- przeżył (152)
- zmarł (266)

7 atrybutów

1. numer klasy pasażera [1, 2, 3]
2. płeć pasażera [male, female]
3. wiek pasażera, atrybut ciągły
4. liczba rodzeństwa/mężów/żon na pokładzie, atrybut ciągły
5. liczba rodziców/dzieci na pokładzie, atrybut ciągły
6. koszt biletu, atrybut ciągły
7. miejsce wejścia na pokład [S, C, Q]

Binarny zbiór danych do przewidywania, czy paczka zostanie dostarczona, czy zwrócona.
https://www.kaggle.com/datasets/pranalibose/amazon-seller-order-status-prediction

Klasy (171)

- zwrot (11)
- dostarczenie (160)

8 atrybutów

1. Data zamówienia
2. Miasto do dostawy, atrybut dyskretny
3. Stan do dostawy, atrybut dyskretny
4. Opis produktu, atrybut dyskretny
5. Liczba produktów, atrybut ciągły
6. Koszt zamówienia, atrybut ciągły
7. Opłata transportowa, atrybut ciągły
8. Sposób płatności [Cash on Delivery, NULL]

Zbiór danych do przewidywania do jakiego gatunku należy wino, na podstawie jego właściwości chemicznych.
https://archive.ics.uci.edu/ml/datasets/wine

Klasy 3 (178)

- 1 (59)
- 2 (71)
- 3 (48)

13 atrybutów, wszystkie mają charakter ciągły

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline