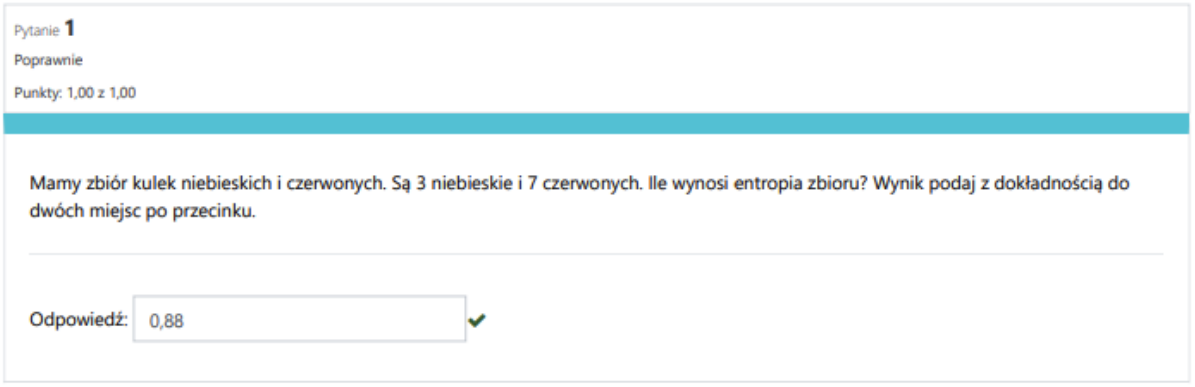
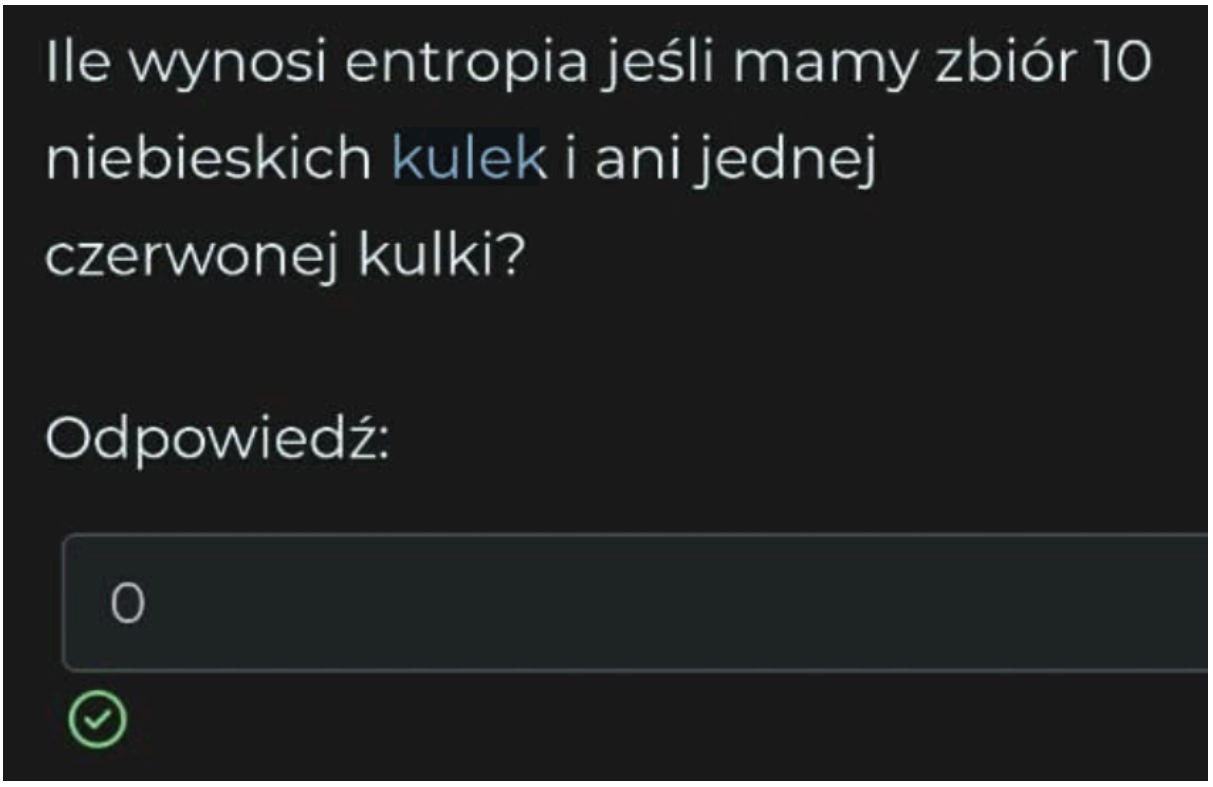
1. Oblicz entropię, entropię warunkową, information gain, intrinsic information, gain ratio.





Pytanie **17**

Zakończone

Punkty maks.: 5,00

Dla zadanych danych oblicz entropię, entropię warunkową, information gain, intrinsic information, gain ratio dla atrybutu **app popularity**.

id	app popularity	features	usability	decision
1	high	5	low	no
2	low	7	high	yes
3	medium	7	medium	yes
4	high	3	low	no
5	low	7	high	yes
6	medium	3	medium	no
7	low	7	high	yes
8	medium	5	medium	yes
9	high	7	low	no
10	medium	5	high	yes

entropia= -4/10 *log2 4/10 - 6/10 * log2 6/10 = 0,966 (te wartości są z atrybutu decyzyjnego)
entropia warunkowa=
high= -0/3*log2 0/3 - 3/3*log2 3/3 = 0
low= -0/3*log2 0/3 - 3/3*log2 3/3 = 0
med= -1/4* log2 1/4 - 3/4*log2 3/4 = 0,81
entropia warunkowa= 0*3/10 + 0*3/10 + 0,81*4/10 = 0,324 (te wartości są z high*ilosc high/10 + low*ilosc low/10 + med*ilosc med/10)
information gain= entropia -entropia warunkowa = 0,642
instrinsic information= -3/10 *log2 3/10 - 3/10 *log2 3/10 - 4/10*log2 4/10 = 1,566 (te wartości to -ilosc wystąpień high/10 * log2 ilosc wystąpień high/10 -ilosc wystąpień low/10 * log2 ilosc wystąpień low/10 -ilosc wystąpień med/10 * log2 ilosc wystąpień med/10)
gain ratio = information gain/ instrinsic information = 0,642/1,566 =0,41

Dla zadanych danych oblicz entropię, entropię warunkową, information gain, intrinsic information, gain ratio dla atrybutu **hotel rating**.

id	hotel rating	rooms	location	decision
1	high	50	downtown	yes
2	medium	30	suburb	no
3	low	10	downtown	yes
4	high	10	countryside	no
5	medium	50	downtown	yes
6	low	10	suburb	no
7	high	30	downtown	yes
8	low	30	countryside	yes
9	medium	50	suburb	no
10	high	50	countryside	yes

Tekst odpowiedzi Pytanie 17

H(S) = -(6/10 log2 (6/10) + 4/10 log2 (4/10)) = 0,971

H(S|H=h) = -(3/10 log2 3/10 + 1/10 * log2 1/10 + 2/10 log2 2/10) = 1.3177

IG(S) = H(S) - H(S|H) = 0.971 - 1.3177 = -0.3467

II = -(3/4 log2 3/4 + 1/3 log2 1/3 + 2/3 log2 2/3) = 1.0566

GR = IG/II = -0.3467/1.0566 = 0.3281

Dla zadanych danych oblicz entropię, entropię warunkową, information gain, intrinsic information, gain ratio dla atrybutu **location**.

id	hotel rating	rooms	location	decision
1	high	50	downtown	yes
2	medium	30	suburb	no
3	low	10	downtown	yes
4	high	10	countryside	no
5	medium	50	downtown	yes
6	low	10	suburb	no
7	high	30	downtown	yes
8	low	30	countryside	yes
9	medium	50	suburb	no
10	high	50	countryside	yes

Tekst odpowiedzi Pytanie 17

entropia(S) = -(6/10)*log2(6/10)-(4/10)*log2(4/10) = 0.97

downtown: y, y, y, y
suburb: n, n, n
countryside: n, y, y

entropia(location=downtown) = 0
entropia(location=suburb) = 0
entropia(location=countryside) = -(1/3)*log2(1/3)-(2/3)*log2(2/3) = 0.92

entropia(S|location) = (4/10)*0+(3/10)*0+(3/10)*0.92 = 0.28

IG(S,location) = entropia(S)-entropia(S|location) = 0.97-0.28 = 0.69

II(S,location) = -(4/10)*log2(4/10)-(3/10)*log2(3/10)-(3/10)*log2(3/10) = 1.57

GR(S,location) = IG(S,location)/II(S,location) = 0.69/1.57 = 0.44

Pytanie **12**

Poprawnie

Punkty: 1,00 z 1,00

Przyrost informacji (ang. info gain) jest zdefiniowany jako (A --- zbiór przykładów uczących, s --- atrybut użyty do podziału w węźle drzewa; Ent symbol obliczenia entropii informacji):

Wybierz jedną odpowiedź:

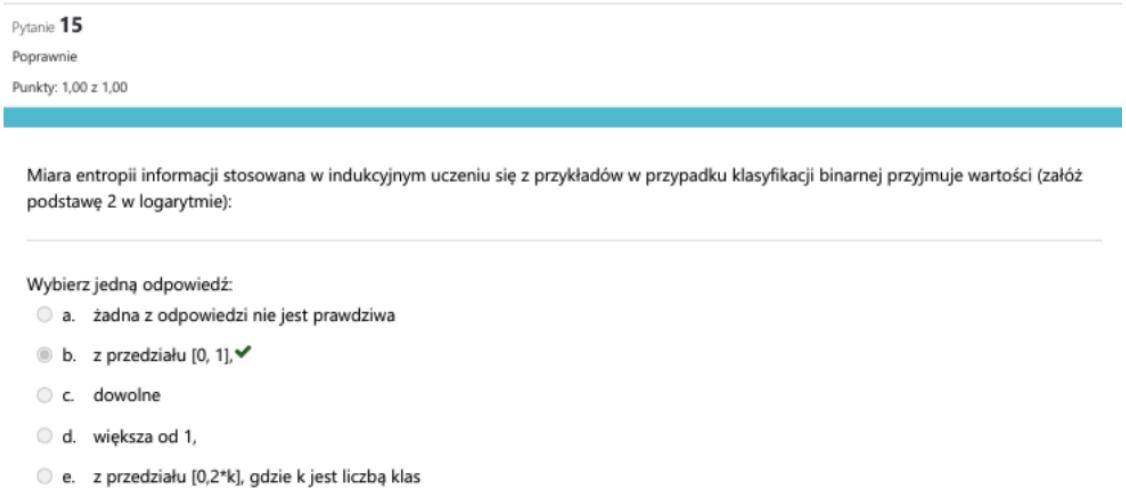
a. Gain(A,s)=Ent(s)-Ent(A|s);

b. Gain(A,s)=Ent(A)-Ent(A|s) ✔

c. Gain(A,s)=Ent(A|s)-Ent(A)

d. Gain(A,s)=Ent(s)-Ent(s|A)

Poprawna odpowiedź to: Gain(A,s)=Ent(A)-Ent(A|s)

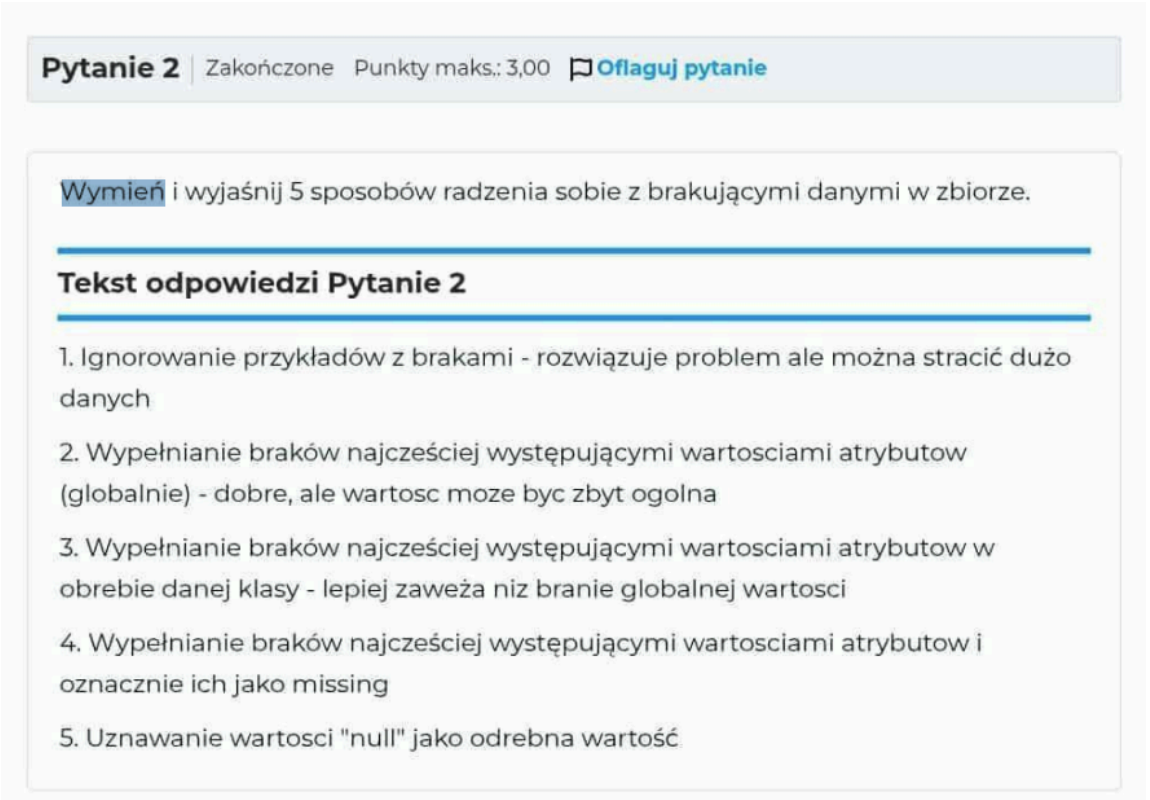


2. Na podstawie powyższych wartości określ, który atrybut będzie najlepszy do wykonania podziału.
BRAK

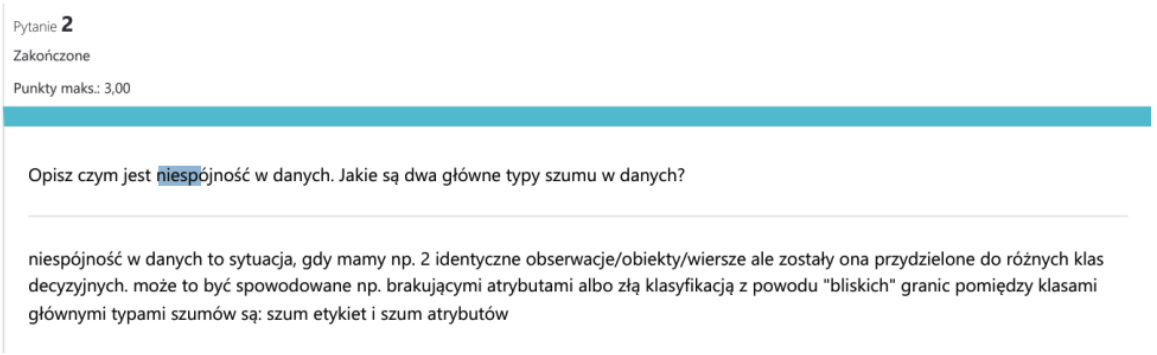
3. Potencjalne problemy z information gain i gain ratio.
BRAK

4. W jaki sposób należy radzić sobie z wartościami ciągłymi na atrybutach?
BRAK

5. W jaki sposób radzić sobie z brakującymi wartościami w zbiorze danych?



6. Źródła niespójności w danych.



Pytanie **2**

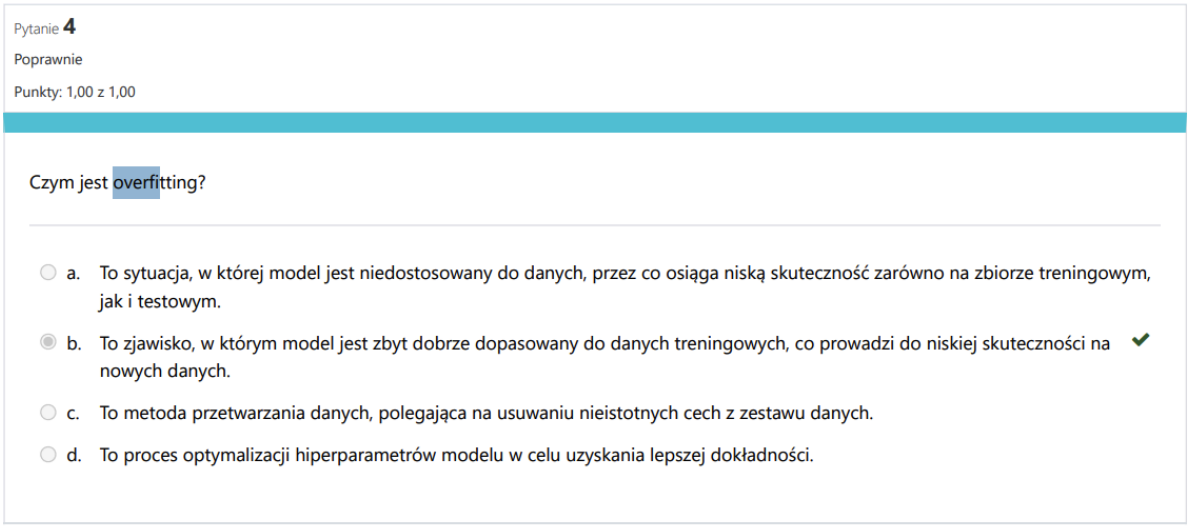
Zakończone

Punkty maks.: 3,00

Opisz czym jest niespójność w danych. Jakie są dwa główne typy szumu w danych?

niespójność w danych to sytuacja, gdy mamy np. 2 identyczne obserwacje/obiekty/wiersze ale zostały ona przydzielone do różnych klas decyzyjnych. może to być spowodowane np. brakującymi atrybutami albo złą klasyfikacją z powodu "bliskich" granic pomiędzy klasami głównymi typami szumów są: szum etykiet i szum atrybutów

7. Przyczyny overfittingu.



Pytanie **4**

Poprawnie

Punkty: 1,00 z 1,00

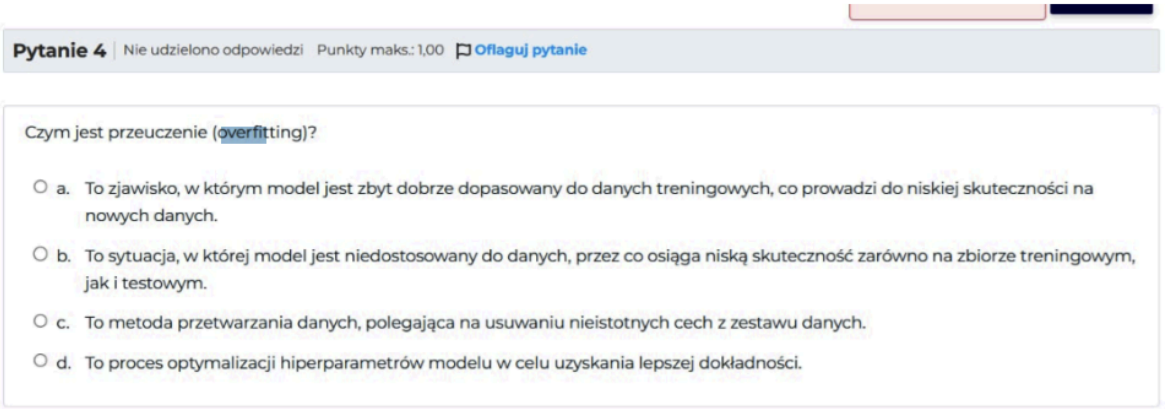
Czym jest overfitting?

a. To sytuacja, w której model jest niedostosowany do danych, przez co osiąga niską skuteczność zarówno na zbiorze treningowym, jak i testowym.

b. To zjawisko, w którym model jest zbyt dobrze dopasowany do danych treningowych, co prowadzi do niskiej skuteczności na nowych danych. ✔

c. To metoda przetwarzania danych, polegająca na usuwaniu nieistotnych cech z zestawu danych.

d. To proces optymalizacji hiperparametrów modelu w celu uzyskania lepszej dokładności.



Pytanie 4 | Nie udzielono odpowiedzi Punkty maks.: 1,00 Oflaguj pytanie

Czym jest przeuczenie (overfitting)?

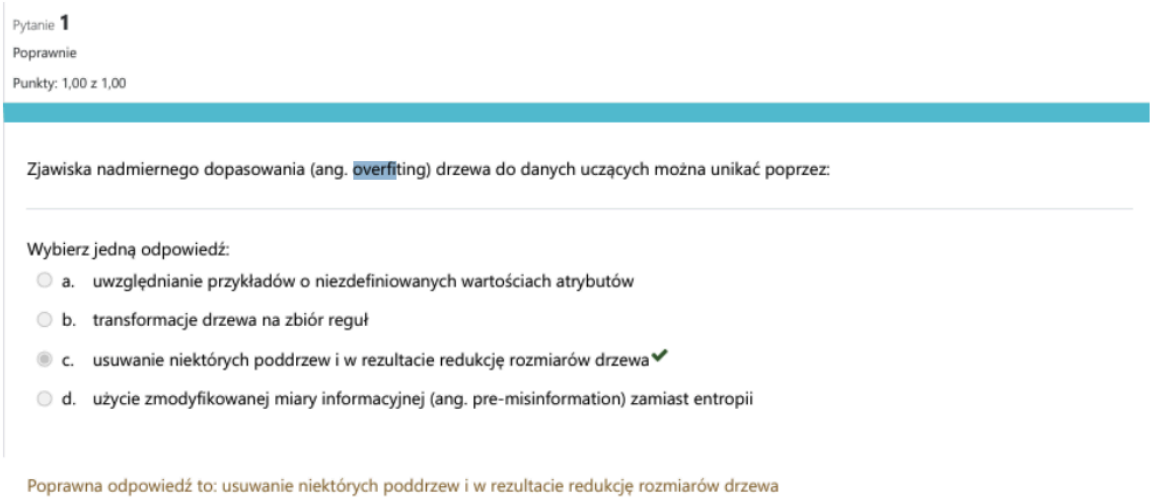
a. To zjawisko, w którym model jest zbyt dobrze dopasowany do danych treningowych, co prowadzi do niskiej skuteczności na nowych danych.

b. To sytuacja, w której model jest niedostosowany do danych, przez co osiąga niską skuteczność zarówno na zbiorze treningowym, jak i testowym.

c. To metoda przetwarzania danych, polegająca na usuwaniu nieistotnych cech z zestawu danych.

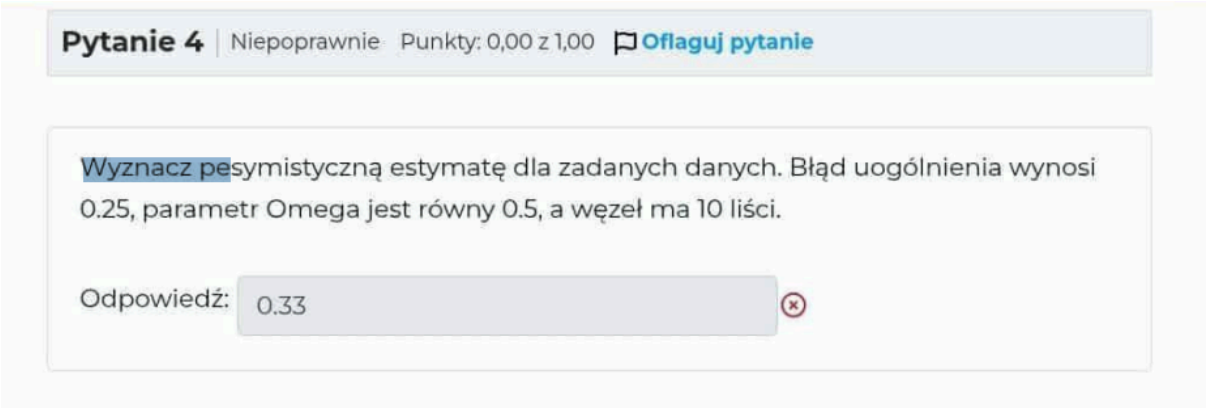
d. To proces optymalizacji hiperparametrów modelu w celu uzyskania lepszej dokładności.

brak odp A

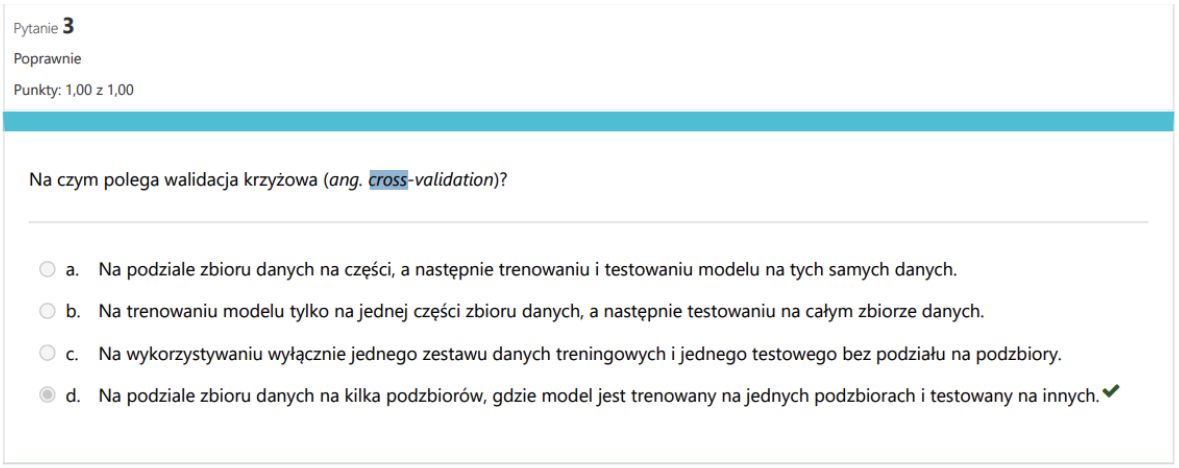


8. Różnica pomiędzy training error i generalization error.
BRAK

9. Wyznaczenie optymistycznej i pesymistycznej estymaty przy błędzie uogólnienia.



10. Na czym polega cross-validation?


Pytanie 3

Poprawnie

Punkty: 1,00 z 1,00

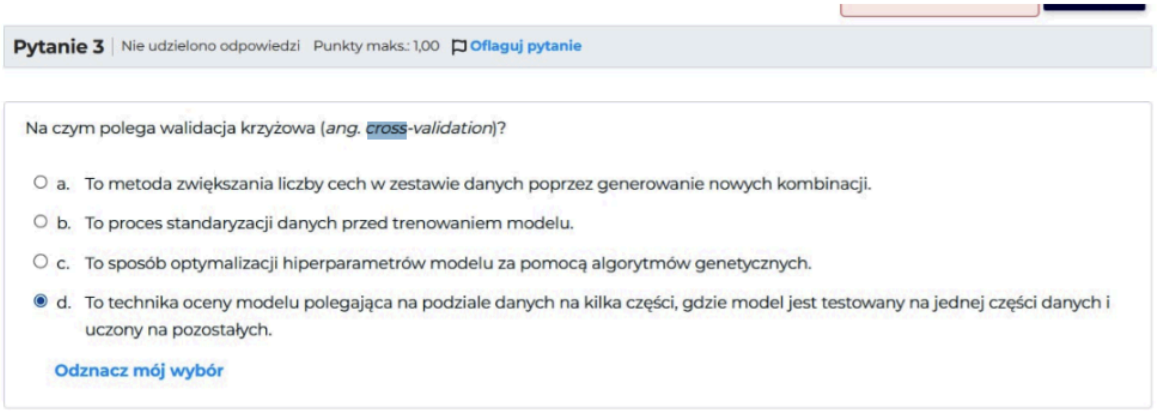
Na czym polega walidacja krzyżowa (*ang. cross-validation*)?

a. Na podziale zbioru danych na części, a następnie trenowaniu i testowaniu modelu na tych samych danych.

b. Na trenowaniu modelu tylko na jednej części zbioru danych, a następnie testowaniu na całym zbiorze danych.

c. Na wykorzystywaniu wyłącznie jednego zestawu danych treningowych i jednego testowego bez podziału na podzbiory.

d. Na podziale zbioru danych na kilka podzbiorów, gdzie model jest trenowany na jednych podzbiorach i testowany na innych. ✔


Pytanie 3 | Nie udzielono odpowiedzi Punkty maks.: 1,00 Oflaguj pytanie

Na czym polega walidacja krzyżowa (*ang. cross-validation*)?

a. To metoda zwiększania liczby cech w zestawie danych poprzez generowanie nowych kombinacji.

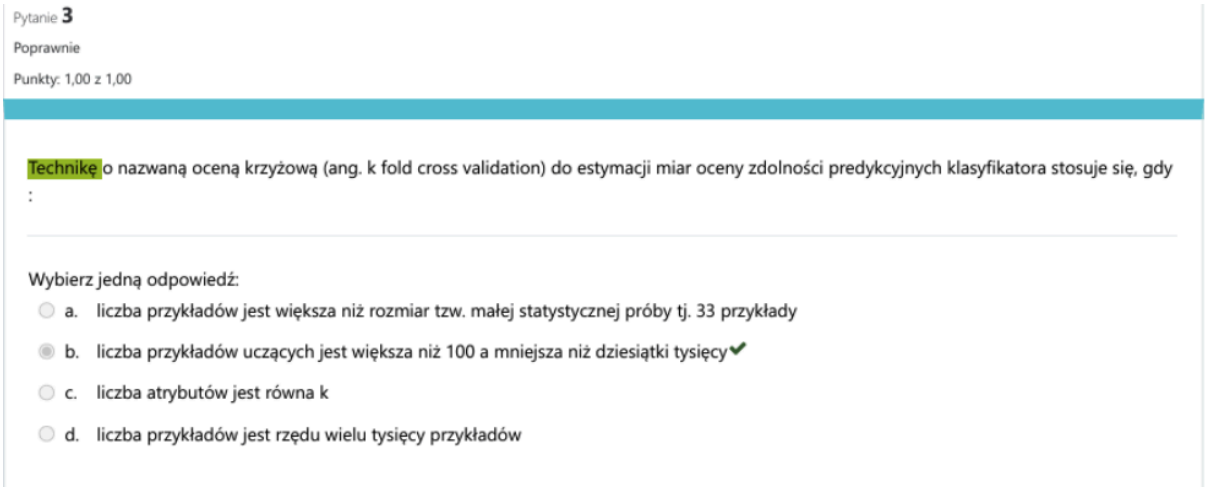
b. To proces standaryzacji danych przed trenowaniem modelu.

c. To sposób optymalizacji hiperparametrów modelu za pomocą algorytmów genetycznych.

d. To technika oceny modelu polegająca na podziale danych na kilka części, gdzie model jest testowany na jednej części danych i uczony na pozostałych.

Odznacz mój wybór

brak odp GIT D


Pytanie 3

Poprawnie

Punkty: 1,00 z 1,00

Technikę o nazwaną oceną krzyżową (ang. k fold cross validation) do estymacji miar oceny zdolności predykcyjnych klasyfikatora stosuje się, gdy :

Wybierz jedną odpowiedź:

a. liczba przykładów jest większa niż rozmiar tzw. małej statystycznej próby tj. 33 przykłady

b. liczba przykładów uczących jest większa niż 100 a mniejsza niż dziesiątki tysięcy ✔

c. liczba atrybutów jest równa k

d. liczba przykładów jest rzędu wielu tysięcy przykładów

Pytanie **11**

Poprawnie

Punkty: 1,00 z 1,00

Ile razy klasyfikowany jest każdy przykład (jako testowy przez klasyfikator) w procedurze oceny krzyżowej (ang. k fold cross validation)?

Wybierz jedną odpowiedź:

a. Dokładnie jeden raz. ✔

b. Przynajmniej raz i nie więcej niż *n*– 1, gdzie *n*to liczba części na jakie jest podzielony zbiór uczący.

c. Jest to zależne od liczby części (ang. podziałów), na jakie został losowo podzielony zbiór uczący?

d. metoda oceny jest techniką wielu losowych podziałów, więc trudno to ustalić

e. *n* razy, gdzie *n* jest zdefiniowane jak w punkcie c.

11. Pre- i post-pruning - na czym polega?

Tzw. „Cost-complexity approach" w indukcji drzew jest wykorzystywane do:

a. Uczenia się oszczędnego z uwagi na zużycie czasu obliczeń

b. Wprowadzenie kosztów użycia poszczególnych atrybutów do korekcji miary entropii

c. Uwzględnienia funkcji kosztu pomyłek w obliczeniach warunków podziału w węzłach

d. Wykonania redukcji rozmiarów drzewa z sumą ważoną kryteriów minimalizacji błędu klasyfikacji oraz rozmiaru drzewa ✔

1. Jak działa algorytm k-nn?

Pytanie **5**

Poprawnie

Punkty: 1,00 z 1,00

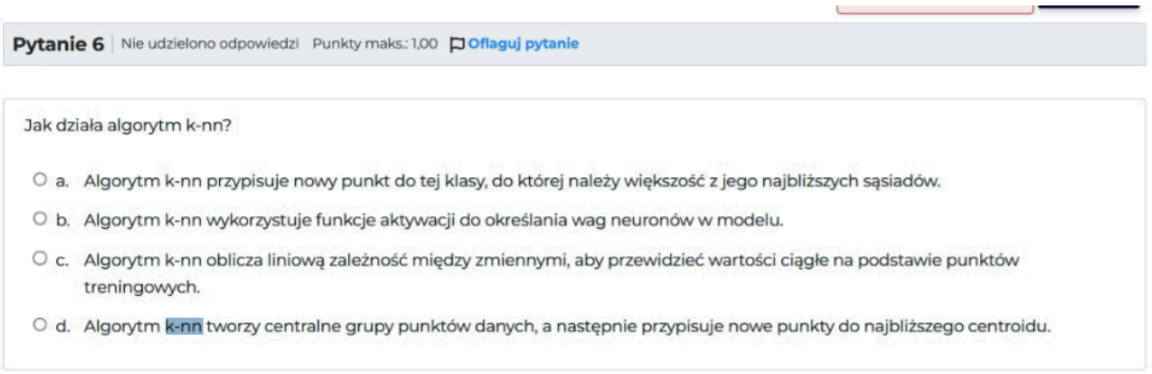
Jak działa algorytm k-nn?

a. Algorytm k-nn klasyfikuje punkty danych za pomocą gradientowego spadku do minimalizacji funkcji kosztu.

b. Algorytm k-nn szuka najbliższych sąsiadów nowego punktu w zbiorze danych, a następnie przypisuje mu klasę na podstawie większości klas wśród sąsiadów. ✔

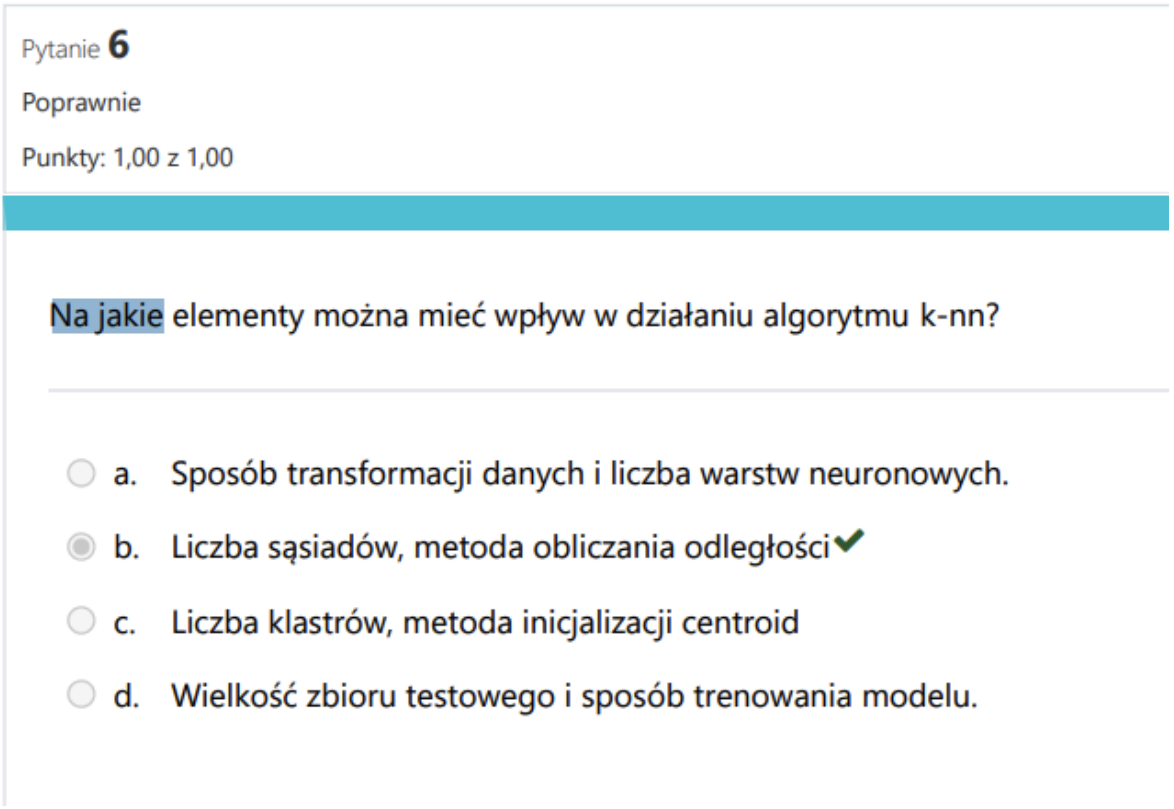
c. Algorytm k-nn tworzy liniowy model, aby przewidywać wartości na podstawie współczynników regresji wyliczonych z danych treningowych.

d. Algorytm k-nn wykonuje podział danych na grupy na podstawie odległości od środkowych punktów (centroidów), aby zminimalizować różnice w obrębie każdej grupy.

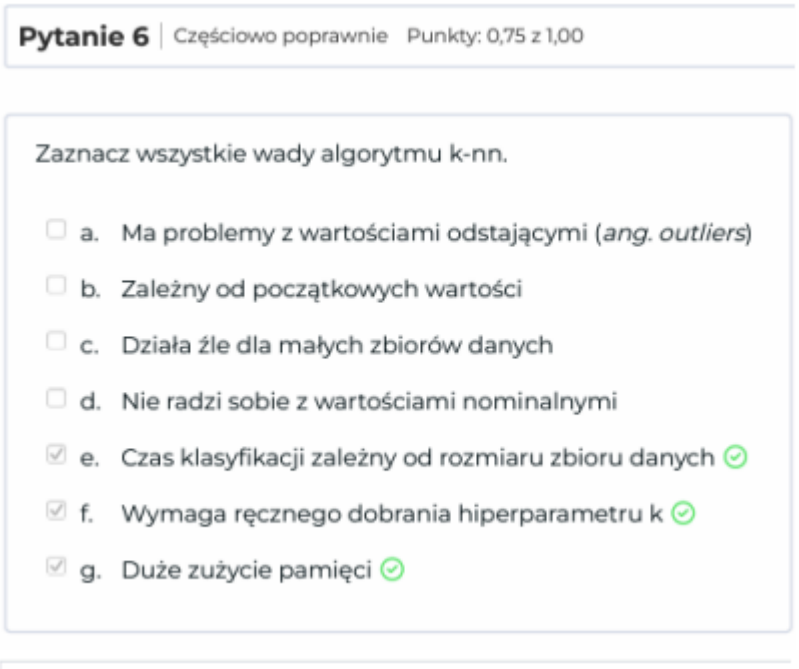


brak odp A

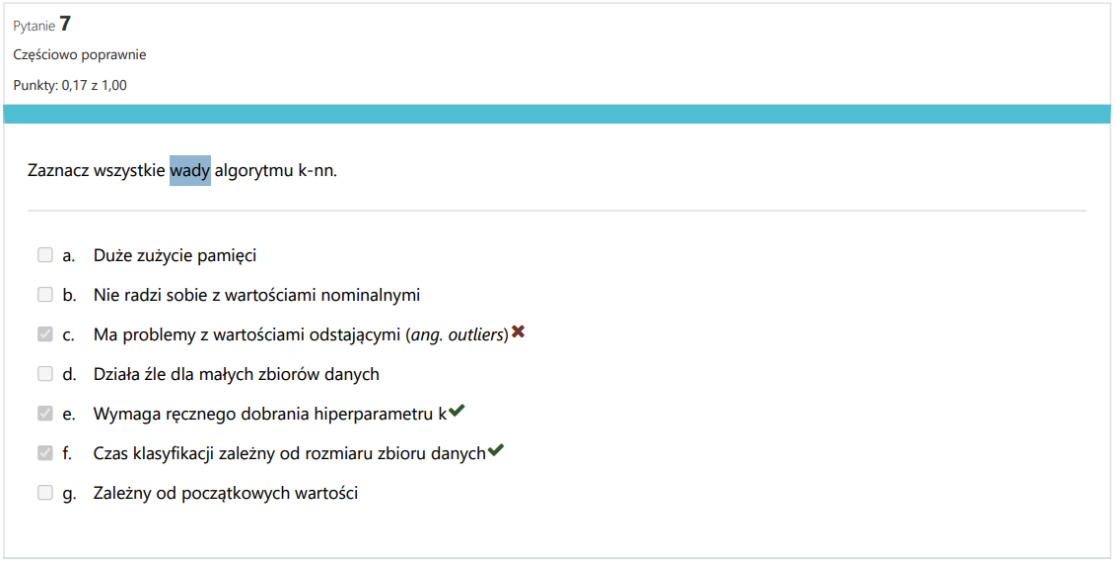
2. Na jakie elementy można mieć wpływ w działaniu algorytmu k-nn?



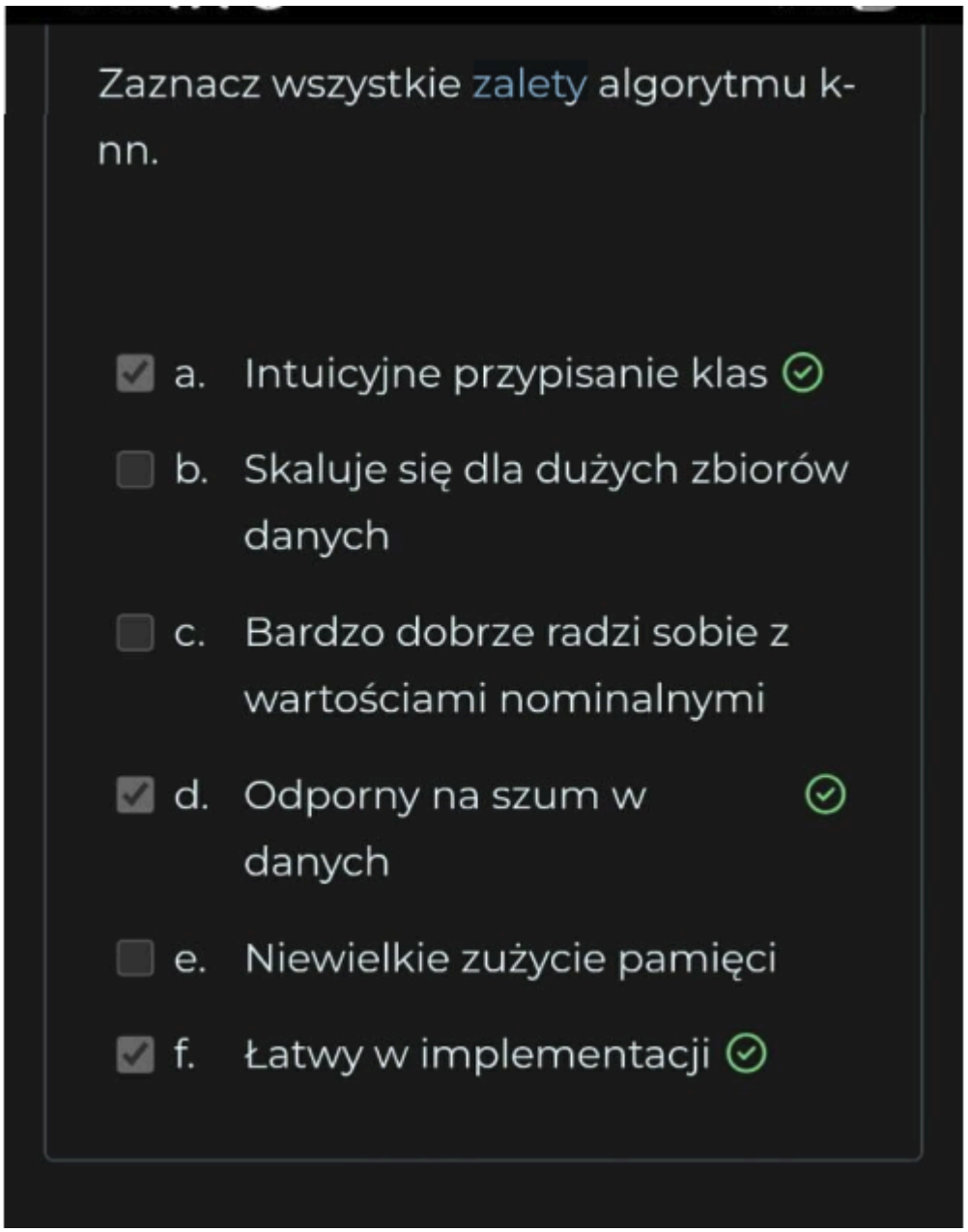
3. Zalety i wady k-nn



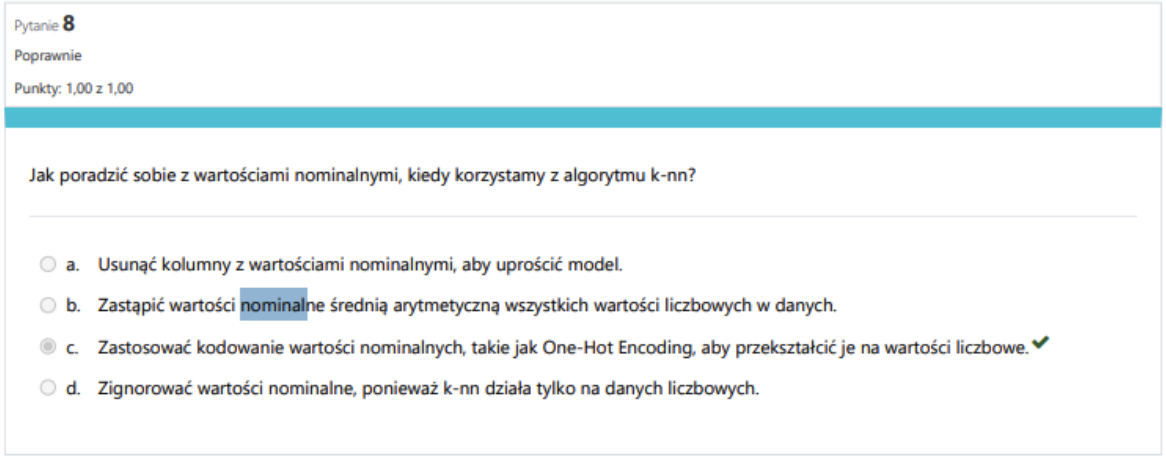
brak odp imo jeszcze dodatkowo D



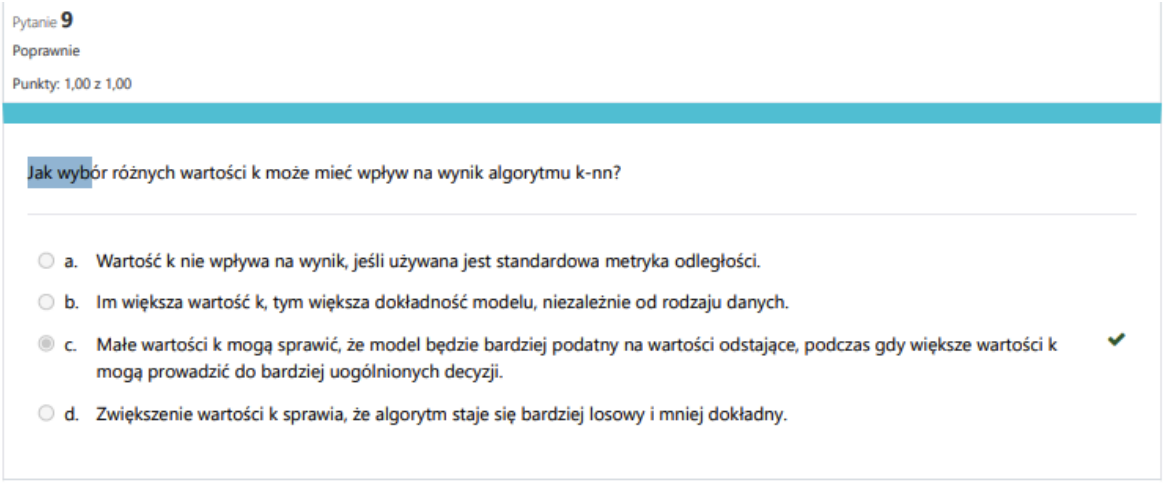
brak odp A, B, E, F

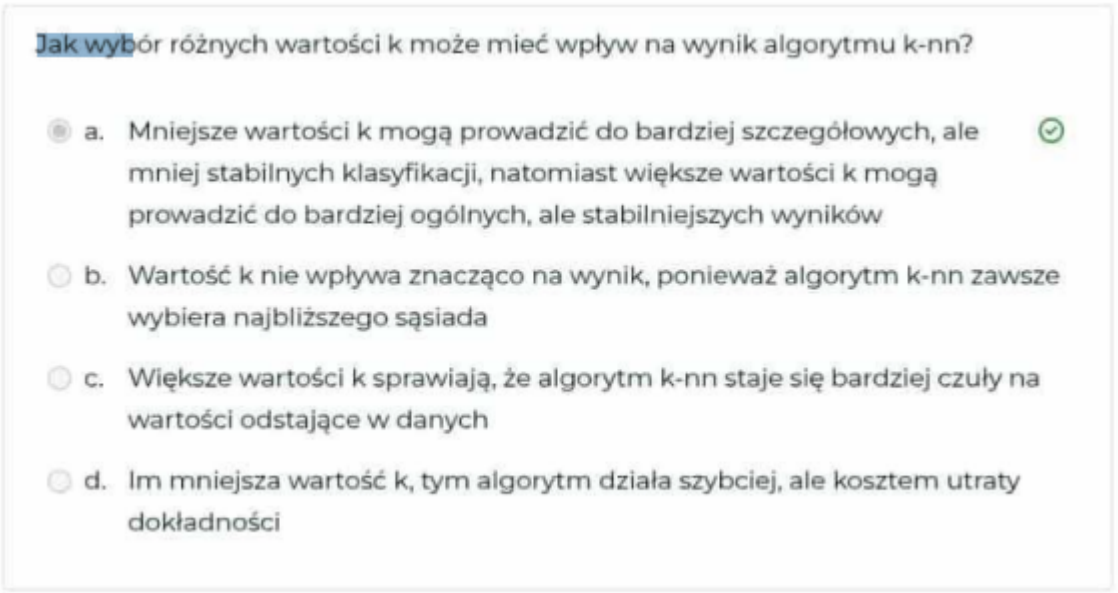


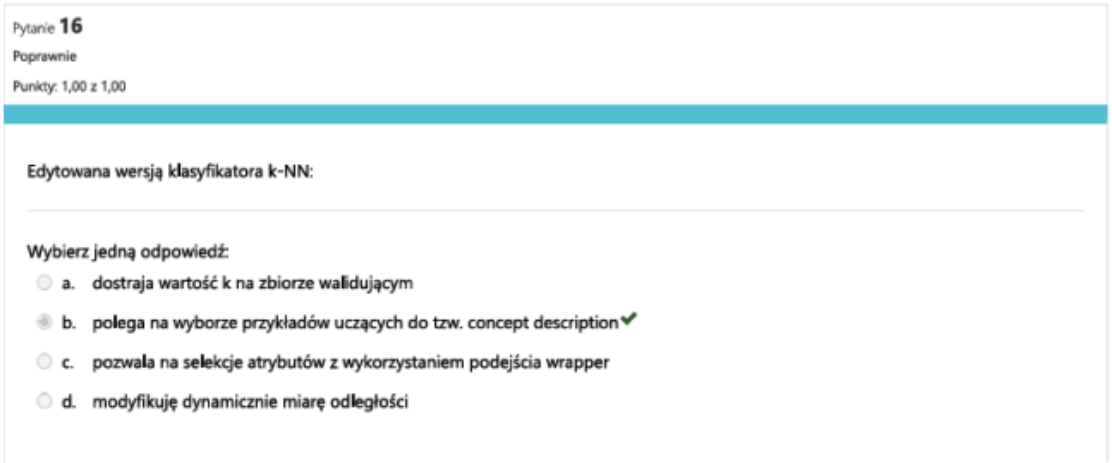
4. Radzenie sobie z wartościami nominalnymi



5. Jak wybór różnych wartości k może mieć wpływ na wynik (k=1, 2, ...,n)







1. Przekleństwo wymiarowości

Pytanie 10 | Poprawnie Punkty: 1,00 z 1,00 Oflaguj pytanie

Czym jest klątwa wielowymiarowości (ang. curse of dimensionality)?

a. To problem wynikający z liniowej zależności między zmiennymi, co powoduje, że model jest mniej dokładny.

b. To zjawisko, w którym wzrost liczby wymiarów danych powoduje, że dane stają się coraz rzadsze, co utrudnia ich efektywną analizę. ✓

c. To technika redukcji liczby zmiennych w zestawie danych w celu optymalizacji działania algorytmu.

d. To sytuacja, w której model traci dokładność z powodu nadmiernego dopasowania do danych treningowych.

Pytanie **11**

Niepoprawnie

Punkty: 0,00 z 1,00

Czym jest klątwa wielowymiarowości (*ang. curse of dimensionality*)?

a. To efekt, w którym złożoność modelu wzrasta w miarę zwiększania liczby warstw w sieci neuronowej. ✗

b. To problem, w którym algorytm k-nn przestaje działać skutecznie, gdy używa się dużych wartości k.

c. To sytuacja, w której zwiększenie liczby wymiarów danych powoduje, że większość punktów znajduje się daleko od siebie, co utrudnia analizę i klasyfikację.

d. To proces automatycznego usuwania korelacji między zmiennymi w celu zwiększenia dokładności modelu.

brak odp C

Pojęcie "przekleństwo wymiarowości" mówi o:

a. wzroście złożoności obliczeniowej wraz ze wzrostem liczby atrybutów

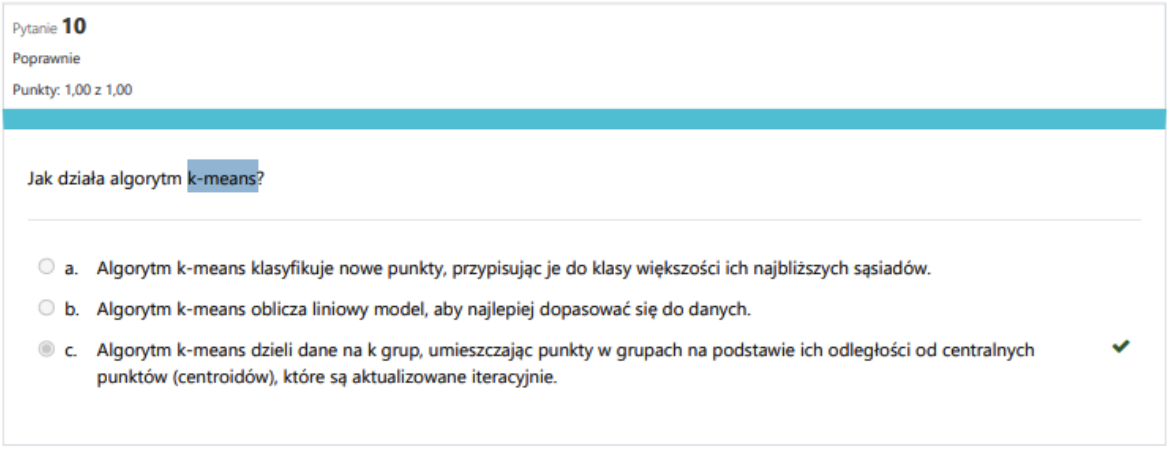
b. żadna z powyższych odpowiedzi nie jest poprawna

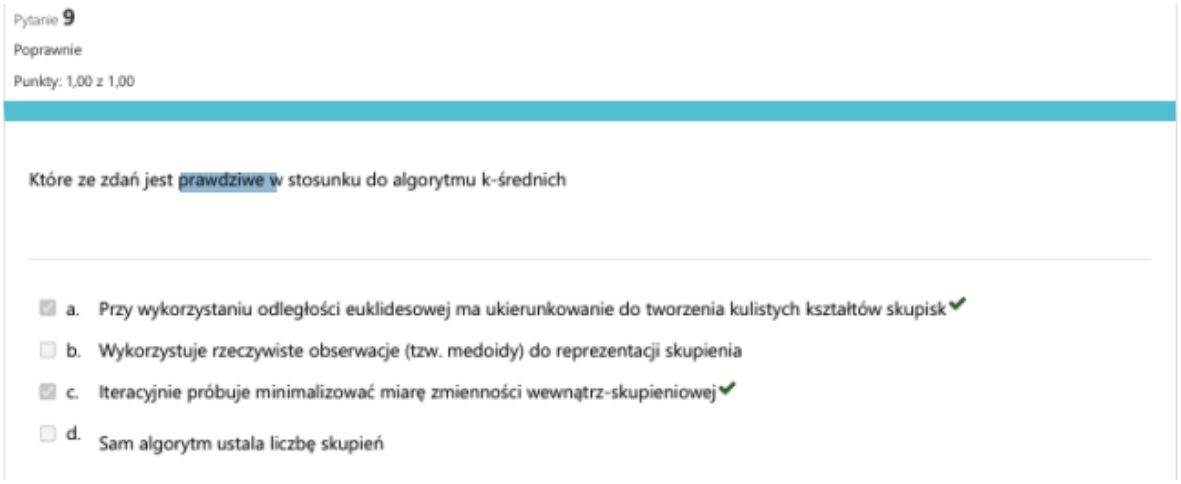
c. konieczności wykładniczego wzrostu liczby przykładów uczących przy rosnącej liczbie atrybutów (aby utrzymać jakość klasyfikacji) ✓

d. spadku jakości klasyfikacji za każdym razem po dodaniu kolejnego atrybutu

3. Czym jest grupowanie?
BRAK

4. Jak działa algorytm?

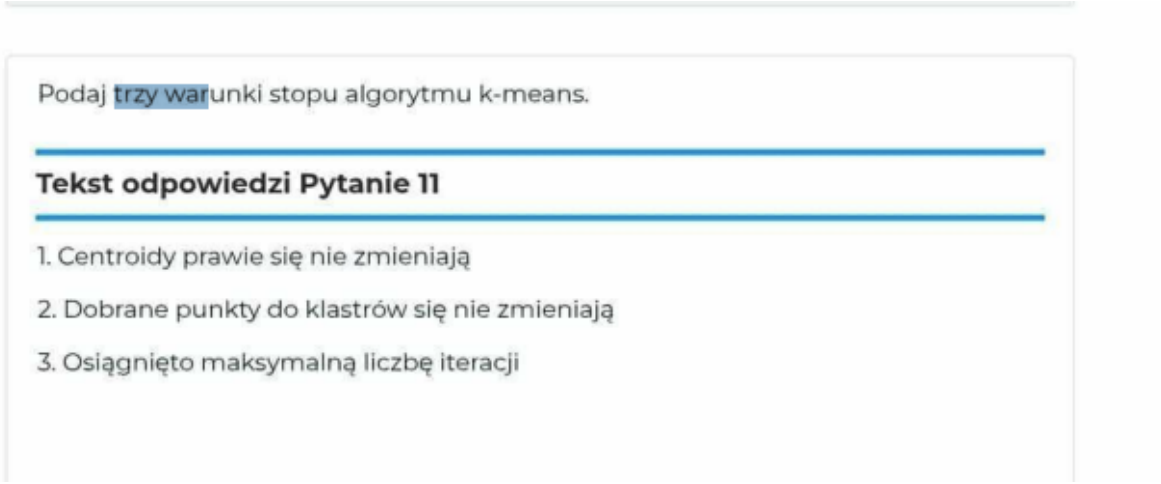




5. Jak wybieramy początkowe centroidy?
BRAK

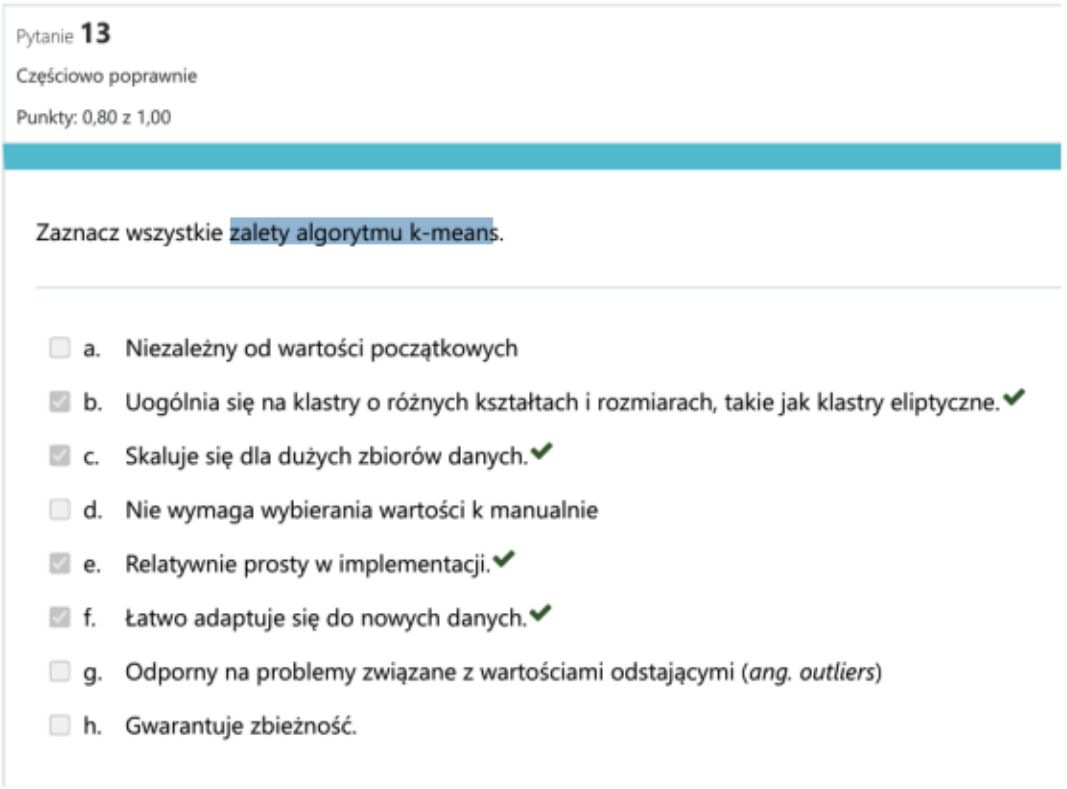
6. Obliczenie jednej iteracji algorytmu
BRAK

7. Jakie są warunki stopu algorytmu?



2. Warunkiem zatrzymania algorytmu k-średnich jest: Wybierz jedną odpowiedź:
   a) zakończenie dyskretyzacji atrybutów liczbowych
   b) utworzenie pełnego drzewa skupień
   c) optymalizacja wartości k - tzn. jest obliczenie
   d) Osiągnięcie stabilizacji zawartości przykładów wewnątrz skupień

8. Zalety i wady algorytmu


Pytanie **13**

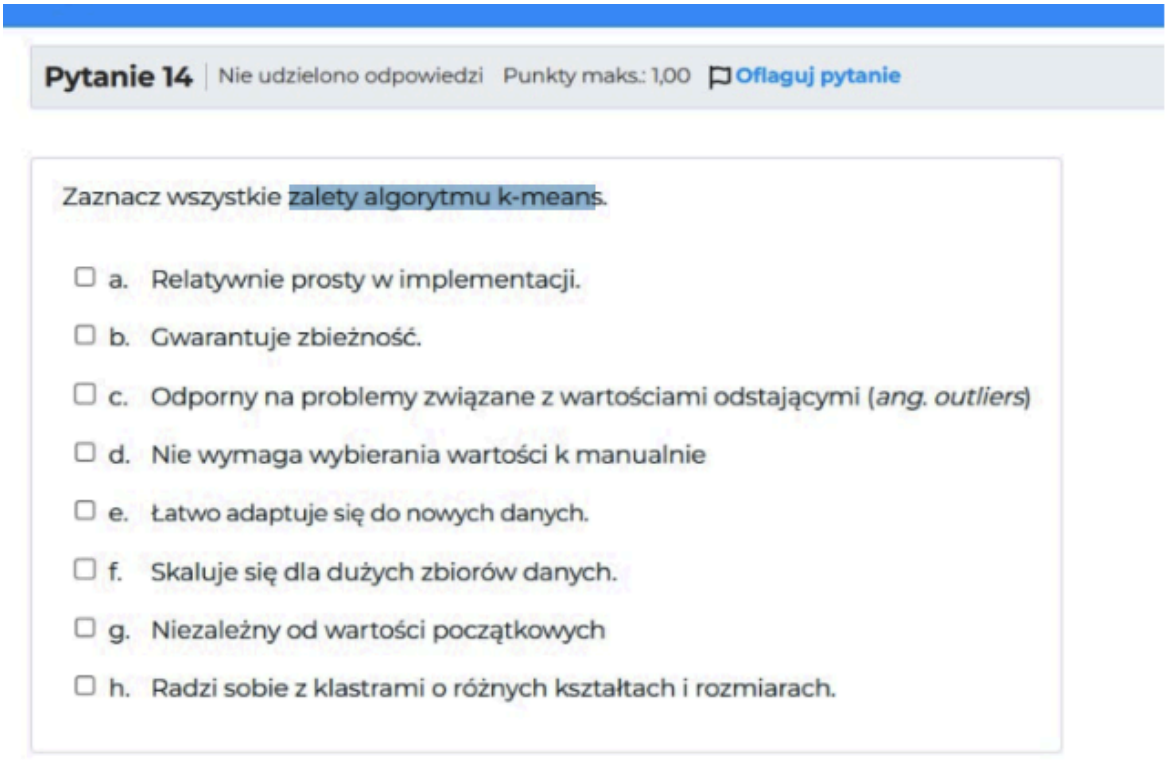
Częściowo poprawnie

Punkty: 0,80 z 1,00

Zaznacz wszystkie zalety algorytmu k-means.

a. Niezależny od wartości początkowych
b. Uogólnia się na klastry o różnych kształtach i rozmiarach, takie jak klastry eliptyczne. ✔
c. Skaluje się dla dużych zbiorów danych. ✔
d. Nie wymaga wybierania wartości k manualnie
e. Relatywnie prosty w implementacji. ✔
f. Łatwo adaptuje się do nowych danych. ✔
g. Odporny na problemy związane z wartościami odstającymi (*ang. outliers*)
h. Gwarantuje zbieżność.

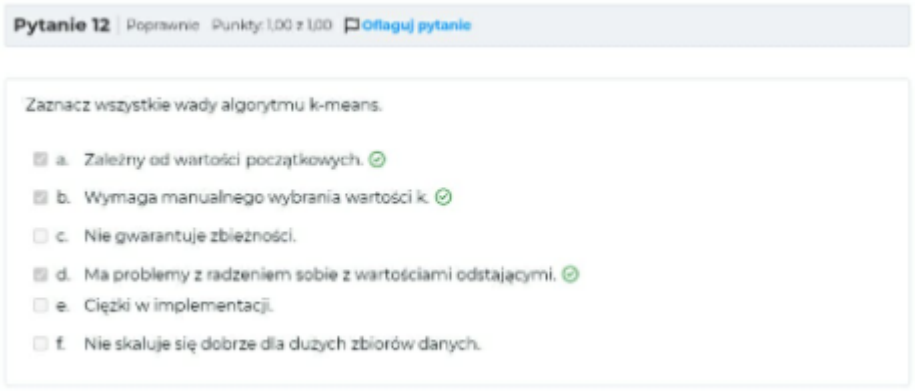
brak odp jeszcze H


Pytanie 14 | Nie udzielono odpowiedzi Punkty maks.: 1,00 Oflaguj pytanie

Zaznacz wszystkie zalety algorytmu k-means.

a. Relatywnie prosty w implementacji.
b. Gwarantuje zbieżność.
c. Odporny na problemy związane z wartościami odstającymi (*ang. outliers*)
d. Nie wymaga wybierania wartości k manualnie
e. Łatwo adaptuje się do nowych danych.
f. Skaluje się dla dużych zbiorów danych.
g. Niezależny od wartości początkowych
h. Radzi sobie z klastrami o różnych kształtach i rozmiarach.

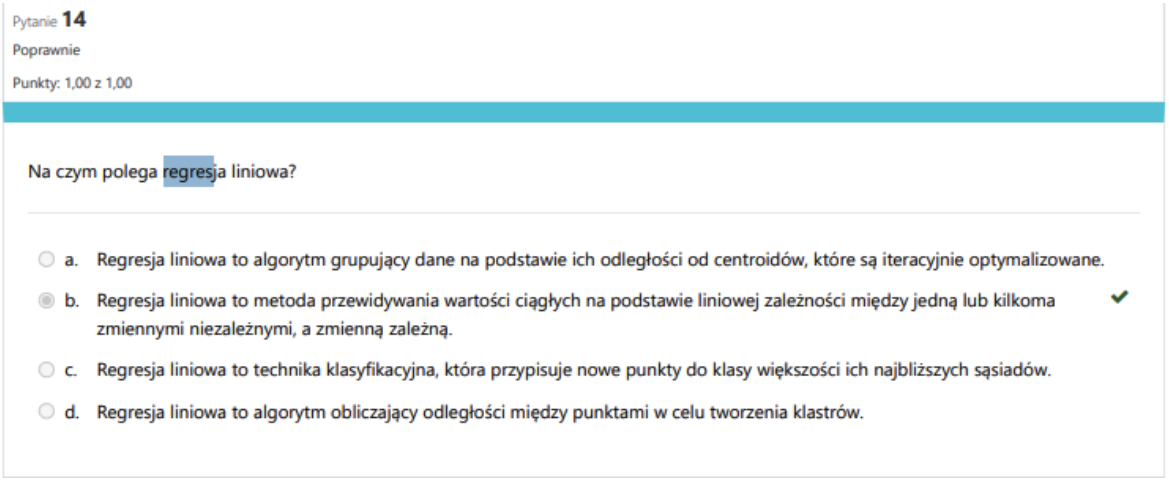
brak odp A, B, E, F

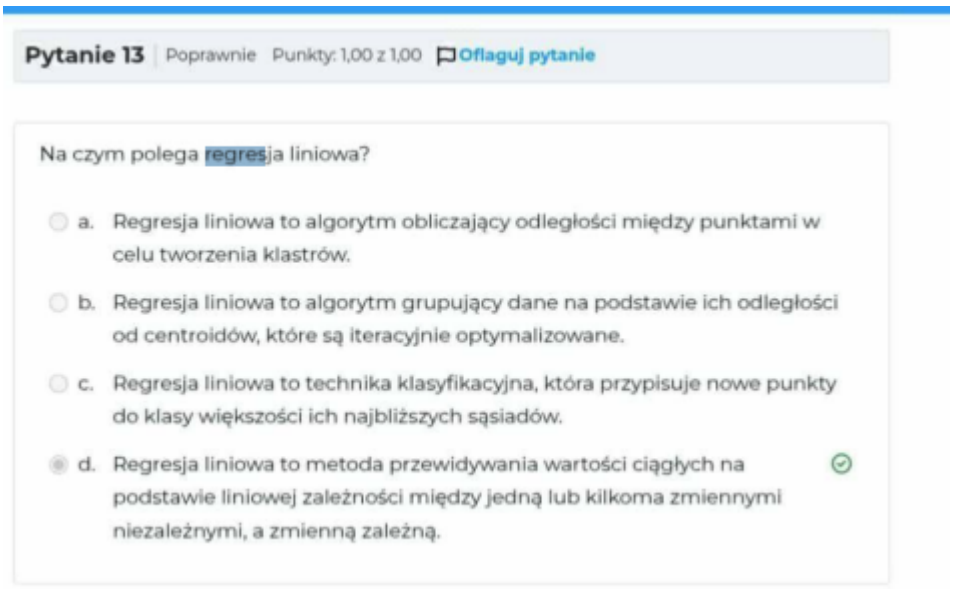


9. Selekcja atrybutów w problemach nienadzorowanych

12. W przypadku uczenia nienadzorowanego (np. grupowanie) do selekcji atrybutów NIE można użyć:
    a) usuwania atrybutów skorelowanych
    b) usuwania atrybutów z niską wariancją
    c) tworzenia nowych atrybutów (np. metodą PCA)
    d) usuwania atrybutów z niską wartością information gain

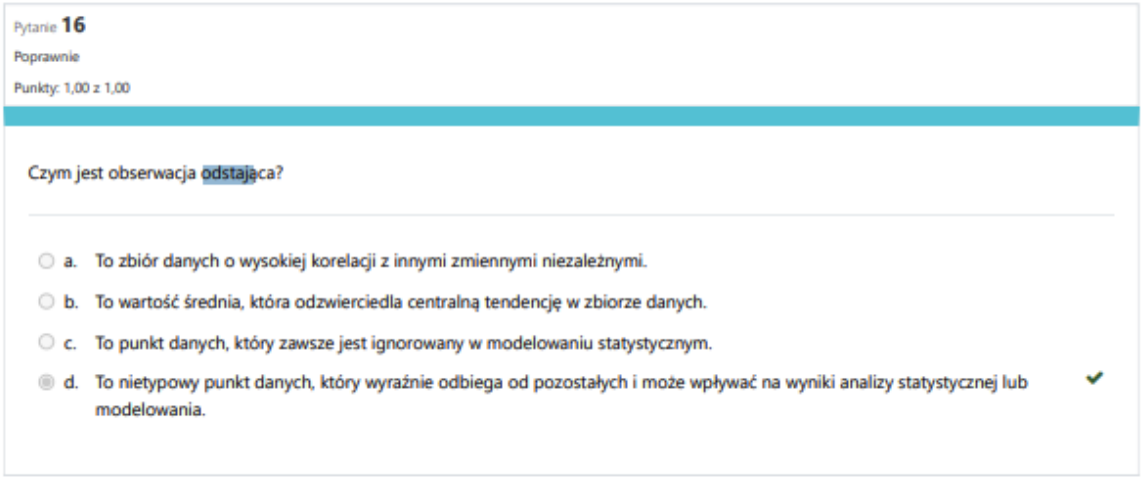
1. Na czym polega regresja liniowa

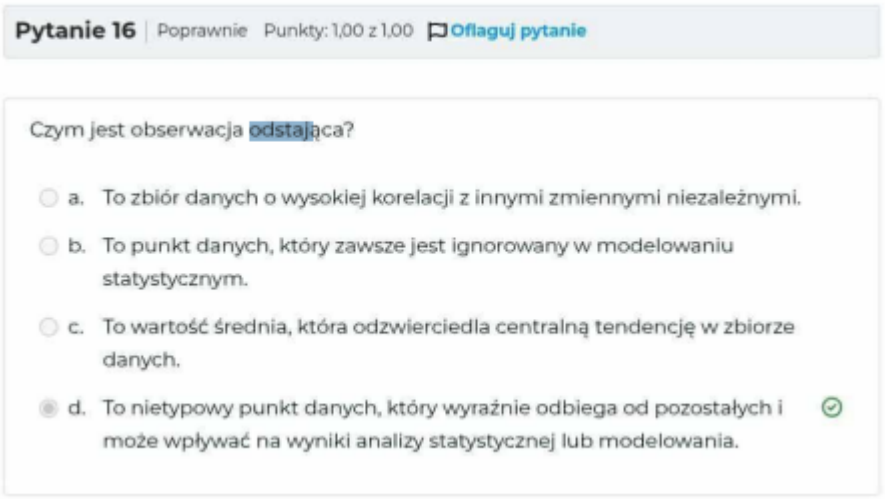




2. Różnica między regresją wielomianową a wieloraką.
BRAK

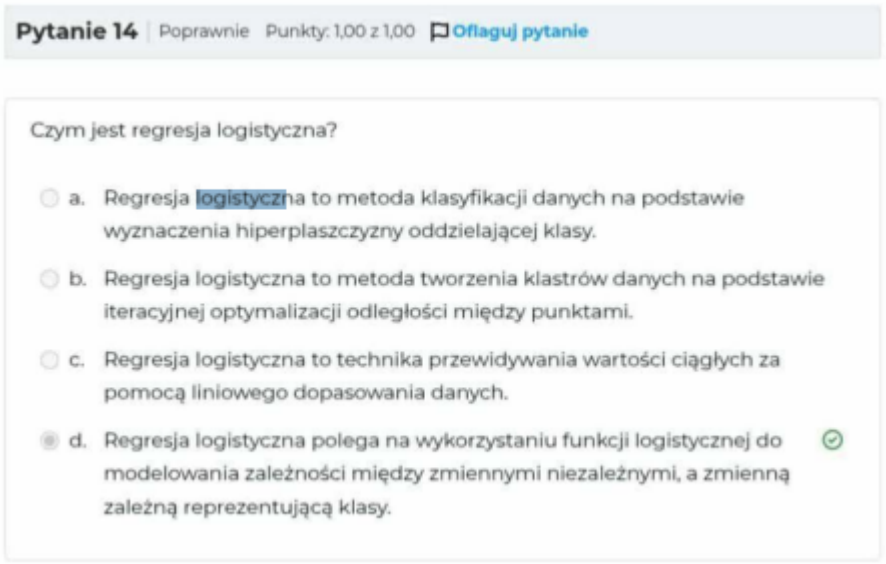
3. Czym jest obserwacja odstająca



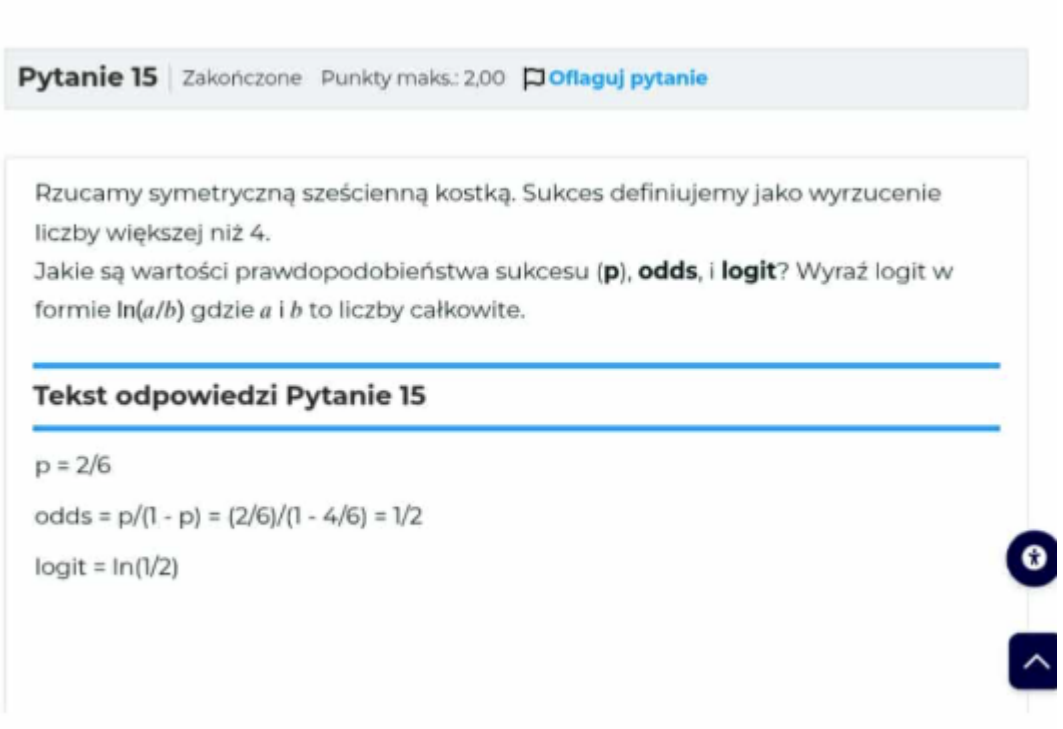


4. W jaki sposób można uznać obserwację za odstającą.
BRAK

5. Czym jest regresja logistyczna



6. Oblicz wartości odds i logit na podstawie danego prawdopodobieństwa.



Pytanie 15 | Zakończone Punkty maks.: 2,00 Oflaguj pytanie

Rzucamy symetryczną sześcienną kostką. Sukces definiujemy jako wyrzucenie liczby większej niż 4.
Jakie są wartości prawdopodobieństwa sukcesu (**p**), **odds**, i **logit**? Wyraź logit w formie ln(*a*/*b*) gdzie *a* i *b* to liczby całkowite.

Tekst odpowiedzi Pytanie 15

p = 2/6

odds = p/(1 - p) = (2/6)/(1 - 4/6) = 1/2

logit = ln(1/2)