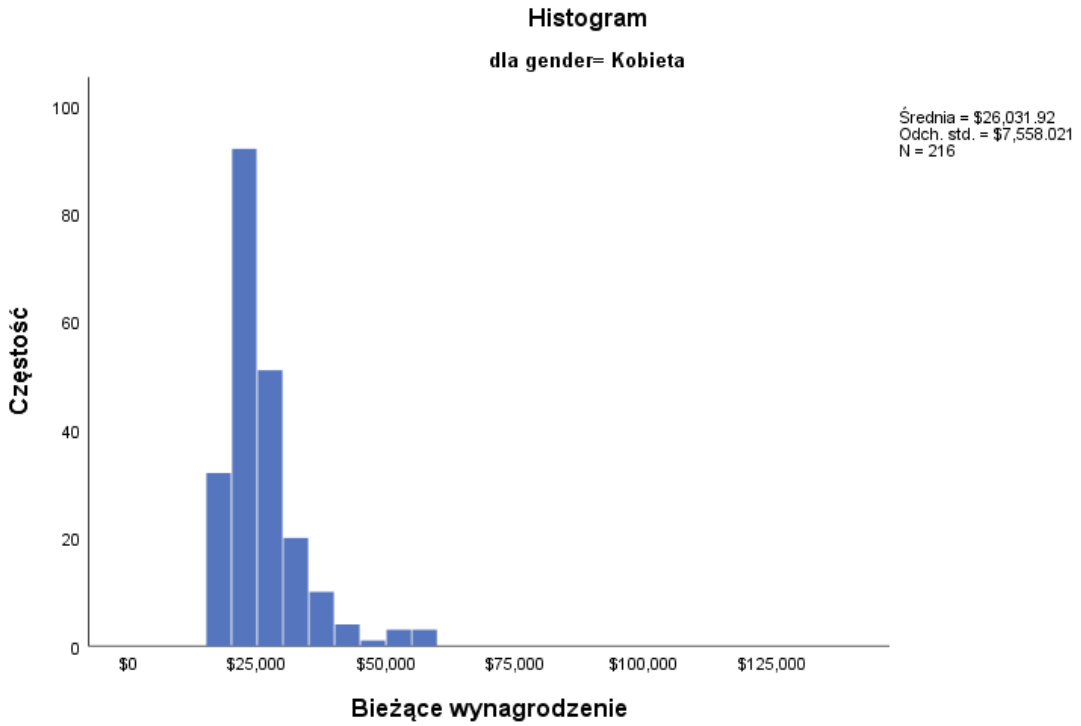
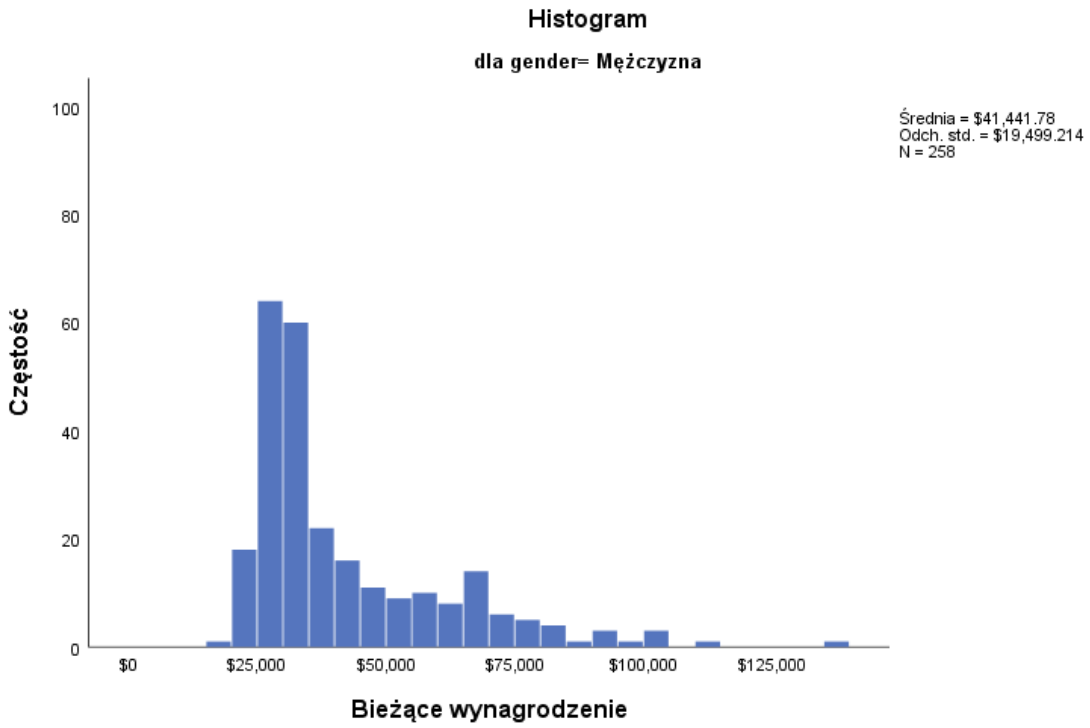
Przykładowy egzamin z Eksploracji danych (Informatyka)

1) Ponieważ nie wszystkie gminy wywiązują się z obowiązku raportowania swojego zadłużenia, władze województwa postanowiły zbudować model, który w oparciu o inne dane będzie dokonywał oceny tego zadłużenia. Wspomniany problem to przykład zadania (zaznacz jedną odpowiedź):
   a) klasyfikacji,
   b) szacowania,
   c) odkrywania reguł,
   d) grupowania.
   UWAGA: Pytanie może opisywać różne sytuacje dotyczące każdego z wymienionych w odpowiedziach zadań eksploracji danych.
2) Zmienna *ryzyko kredytowe* o wartościach: niskie, średnie, wysokie, ma poziom (zaznacz jedną odpowiedź):
   a) nominalny,
   b) porządkowy,
   c) ilościowy (inaczej: skala).
   UWAGA: W pytaniu mogą pojawić się różne zmienne każdego z wymienionych w odpowiedziach poziomów pomiaru.
3) Zmienna *wzrost respondenta w cm* ma rozkład z minimum równym 163, maksimum 208, średnią 176 i odchyleniem standardowym 20. Respondent A ma wzrost 181 cm. Jaką wartość dla respondenta A ma opisana wyżej zmienna po normalizacji min-max? Odpowiedź podaj w postaci ułamka dziesiętnego z dokładnością do 2 miejsc po PRZECINKU.
   UWAGA: Pytanie może dotyczyć także standaryzacji.
4) Histogramy pokazują rozkład wynagrodzenia kobiet i mężczyzn w pewnej firmie.

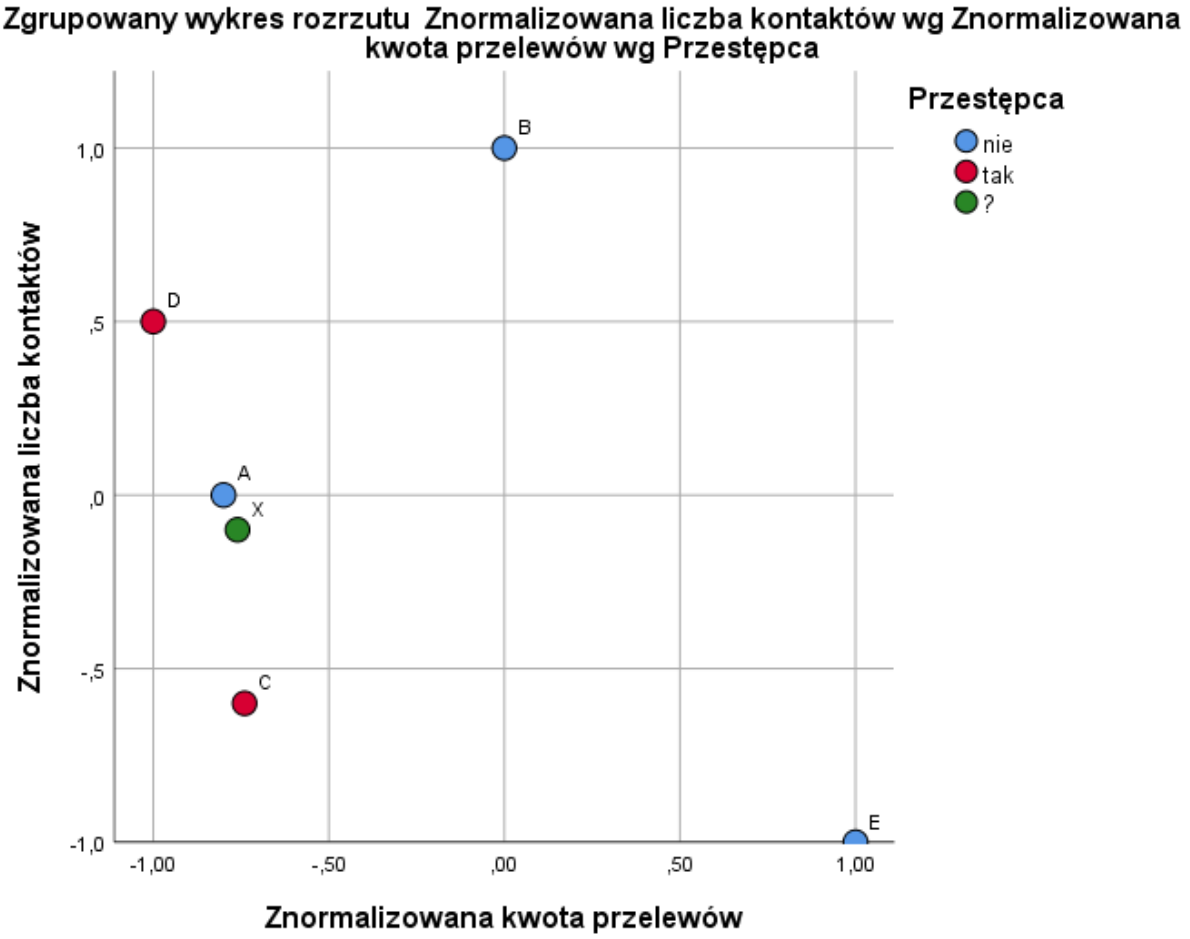




Zaznacz stwierdzenia prawdziwe:

a) W tej firmie kobiety są lepiej wynagradzane niż mężczyźni.
b) Rozkład wynagrodzeń w obu grupach jest rozkładem normalnym.
c) W tej firmie są pojedyncze osoby zarabiające znacznie więcej od pozostałych i są to mężczyźni.
d) Rozkład wynagrodzeń w obu grupach jest prawostronnie skośny.
UWAGA: Należy się spodziewać pytań o histogramy, wykresy skrzynkowe oraz słupkowe różnego typu.

5) W sytuacji przedstawionej na rysunku



algorytm k najbliższych sąsiadów z k=2 i metryką euklidesową (wybierz jedną odpowiedź):

a) określi nową obserwację X jako należącą do klasy tak,
b) określi nową obserwacje X jako należącą do klasy nie,
c) będzie remis,
d) nie wykona się.
UWAGA: Należy się spodziewać w pytaniu wartości k od 1 do 5 i metryki euklidesowej albo miejskiej.

6) Na podstawie podanej niżej macierzy pomyłek dla zmiennej *X* o wartościach 0 i 1, traktując 1 jako wartość przewidywaną, określ w procentach (podaj tylko liczbę bez symbolu %):

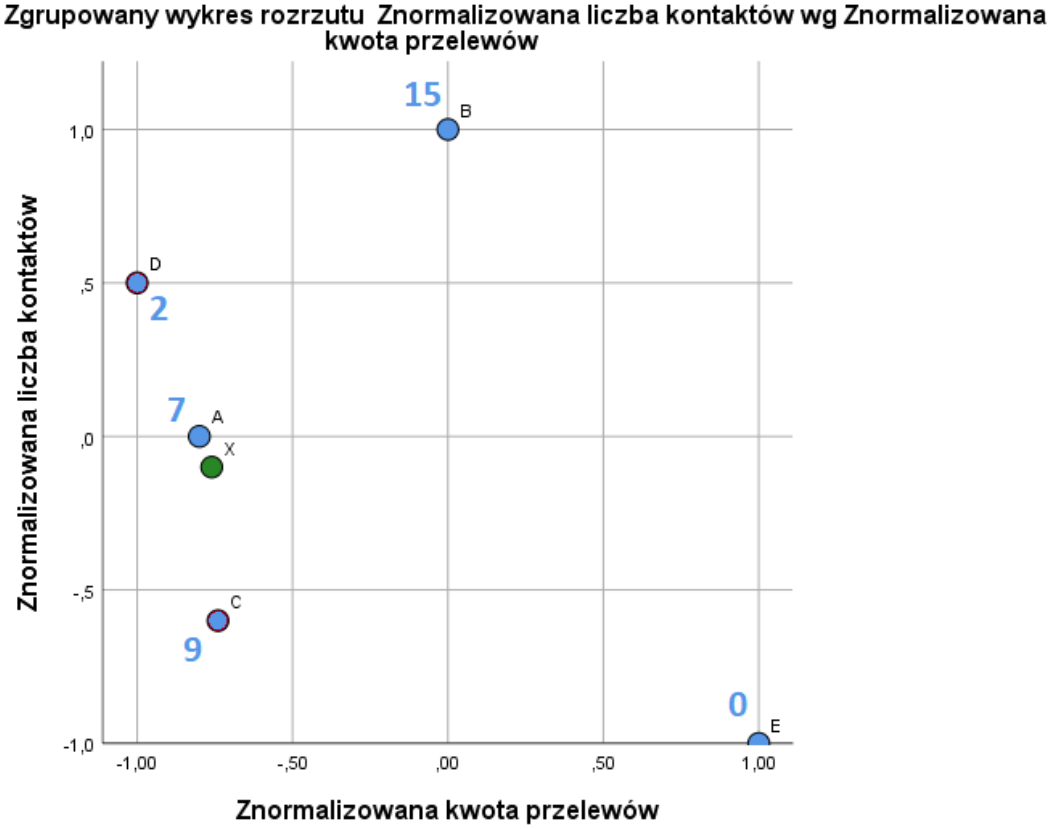
Zmienna *X*		Wartości przewidywane	
		0	1
Wartości obserwowane	0	90	90
	1	80	240

a) Trafność

b) Czułość

c) Specyficzność (swoistość)

7) W sytuacji przedstawionej na rysunku:



algorytm k najbliższych sąsiadów z k=3 i metryką euklidesową opierając się na medianie oszacuje nieznaną wartość dla obserwacji X jako … (podaj odpowiedź, zaokrąglając ją do całości).
UWAGA: Mogą się pojawić różne wartości k oraz średnia lub mediana.

8) Spośród wymienionych algorytmów wybierz te, które mogą być stosowane i w zagadnieniu klasyfikacji, i szacowania.

a) *k* najbliższych sąsiadów,

b) perceptron wielowarstwowy (MLP),

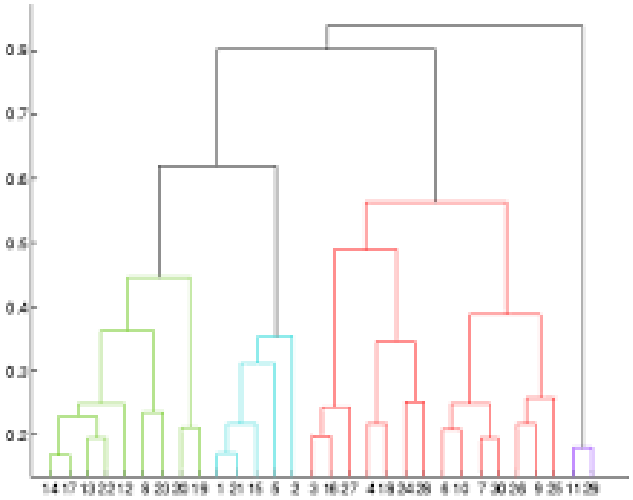
c) regresja liniowa,

d) k średnich.
UWAGA: Należy umieć wybrać algorytmy służące do klasyfikacji, szacowania, grupowania oraz budowy reguł asocjacyjnych.

9) Analizując pewne zagadnienie, otrzymano model regresji określony wzorem y=-7x+21. Oznacza to, że (wybierz jedną odpowiedź):

a) Wraz ze wzrostem wartości zmiennej x o 1 wartość zmiennej y rośnie o 7.
b) Wraz ze wzrostem wartości zmiennej x o 1 wartość zmiennej y rośnie o 21.
c) Wraz ze wzrostem wartości zmiennej x o 1 wartość zmiennej y maleje o 7.
d) Wraz ze wzrostem wartości zmiennej y o 1 wartość zmiennej x maleje o 7.

10) W oparciu o przedstawiony na wykresie dendrogram (proszę się nie sugerować kolorami)



należy przyjąć, że liczba skupień (grup) w zbiorze danych wynosi (wybierz jedną odpowiedź):

a) 2,
b) 3,
c) 4
d) Więcej niż 4.

11) Które z wymienionych cech ma reguła

wiek = [20…30] i chipsy_Lays = 1 → paluszki = 0

a) binarna ,
b) ilościowa,
c) wielowymiarowa,
d) jednopoziomowa.

UWAGA: Należy umieć określić czy reguła jest binarna/ilościowa, jedno-/wielowymiarowa, jedno-/wielopoziomowa.

12) Na 1000 klientów pewnego sklepu 600 kupiło napój gazowany, 180 kupiło sok, z czego 30 kupiło jednocześnie napój gazowany i sok. Dla reguły *Jeśli sok, to napój gazowany* oblicz i podaj wynik z dokładnością do 2 miejsc po przecinku:

a) pokrycie
b) wsparcie
c) ufność
d) wzrost
e) wdrażalność

13) Zaznacz zdania prawdziwe dla modelu sieci neuronowej MLP:

a) Może mieć tylko jedną warstwę wyjściową.

b) W przypadku klasyfikacji liczba neuronów w warstwie wyjściowej jest większa niż w przypadku szacowania.

c) Uczenie sieci polega na dobieraniu odpowiedniej funkcji aktywacji w neuronach.

d) W warstwie wejściowej występuje funkcja aktywacji.

UWAGA: Należy się spodziewać różnych pytań o własności MLP.

14) Które z wymienionych algorytmów są wrażliwe na występowanie obserwacji odstających?

a) k najbliższych sąsiadów,

b) drzewo klasyfikacyjno-regresyjne CRT,

c) perceptron wielowarstwowy,

d) k średnich.

UWAGA: Należy się spodziewać różnych algorytmów do wyboru.

15) Które z wymienionych algorytmów wymagają normalizacji lub standaryzacji zmiennych?

a) drzewo klasyfikacyjno-regresyjne CRT,

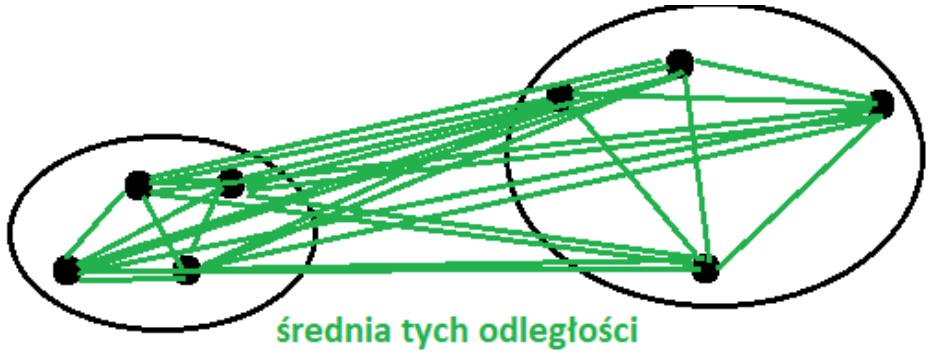
b) perceptron wielowarstwowy,

c) grupowanie hierarchiczne,

d) dwustopniowa analiza skupień

UWAGA: Należy się spodziewać różnych algorytmów do wyboru.

16) Na rysunku poniżej przedstawiono dwa skupienia 4-elementowe. Pojęcie odległości jest rozumiane jako odległość euklidesowa w $\mathbf{R}^2$.



Którą z metod liczenia odległości pomiędzy grupami ilustruje ten rysunek? Wybierz jedną odpowiedź.

a) Pojedynczego połączenia (najbliższego sąsiedztwa),

b) Całkowitego połączenia (najdalszego sąsiedztwa),

c) Średnich grupowych (średniej odległości między skupieniami),

d) Centroidalnego połączenia,

e) Średniego połączenia.

UWAGA: Należy się spodziewać rysunków ilustrujących wszystkie z wymienionych metod liczenia odległości miedzy grupami.

17) Zaznacz zdania prawdziwe dla algorytmu drzew klasyfikacyjno-regresyjnych CRT.

a) Węzeł początkowy drzewa jest nazywany korzeniem.

b) Algorytm w każdym węźle dokonuje podziału, dla którego miara nieczystości jest największa.

c) Drzewo CRT nie zadziała w przypadku wystąpienia braku danych dla zmiennych użytych w węzłach do podziału.

d) Drzewa CRT mogą być stosowane dla ciągłej zmiennej celu.

UWAGA: Należy się spodziewać różnych pytań o własności algorytmu CRT.

18) Których z algorytmów grupowania nie wybierzesz w sytuacji, gdy liczba obserwacji jest duża?

a) k średnich,

b) grupowanie hierarchiczne,

c) dwustopniowa analiza skupień.

UWAGA: Należy się spodziewać opisu różnych sytuacji, do których należy dobrać odpowiedni algorytm.

19) Tabela poniżej jest wynikiem zastosowania algorytmu k średnich dla danych dotyczących gmin. Zmienne zostały wcześniej zestandaryzowane.

	Grupa 1.	Grupa 2.	Grupa 3.
Liczba ludności	-2,33	2,09	-0,11

Dochód	-0,09	1,74	2,03
Wydatki	0,12	-0,09	-2,24

W grupie 2. znalazły się gminy:

a) O małej liczbie ludności, średnim dochodzie i średnich wydatkach.

b) O dużej liczbie ludności, dużym dochodzie i średnich wydatkach.

c) O średniej liczbie ludności, dużym dochodzie i małych wydatkach.

d) O dużej liczbie ludności, dużym dochodzie i małych wydatkach.

20) Który z wymienionych algorytmów do określania klasy nowej obserwacji wykorzystuje mechanizm (ważonego lub nie) głosowania większościowego?

a) Bagging

b) Boosting

c) Las losowy

d) C4.5

UWAGA: Można się spodziewać innych pytań o wymienione algorytmy.

21) Zaznacz zdania prawdziwe dla modelu sieci neuronowej RBF:

a) Może mieć tylko jedną warstwę wyjściową.

b) W przypadku klasyfikacji liczba neuronów w warstwie wyjściowej jest większa niż w przypadku szacowania.

c) Uczenie sieci polega na wielokrotnym korygowaniu wag przypisanych do połączeń pomiędzy neuronami.

d) Współczynniki występujące we wzorach funkcji aktywacji są zależne od umiejscowienia i wielkości skupisk występujących w danych.

UWAGA: Należy się spodziewać różnych pytań o własności RBF.

22) Na podstawie poniższej tabeli

Składowa	Początkowe wartości własne		
	Ogółem	% wariancji	% skumulowany
1	4,398	39,984	39,984
2	1,731	15,741	55,724
3	1,494	13,579	69,304
4	1,165	10,592	79,896
5	,638	5,803	85,699
6	,505	4,591	90,290
7	,468	4,253	94,543
8	,310	2,818	97,361
9	,136	1,240	98,601
10	,085	,770	99,371
11	,069	,629	100,000

wyznacz liczbę składowych głównych zgodnie z kryterium części wariancji wyznaczanej przez składowe główne z progiem 80%.

UWAGA: należy się spodziewać tabeli lub wykresu osypiska i różnych kryteriów.

23)

Na podstawie danych z poniższej tabeli

Lp.	Y obserwowane	Y przewidywane
1	15	18
2	17	16
3	18	19
4	16	15
5	17	21

Oblicz wartość błędu średniokwadratowego modelu użytego do szacowania. Czy średni błąd bezwzględny będzie większy od pierwiastka z wyznaczonej przez Ciebie wartości? Wpisz tak lub nie.

UWAGA: Należy umieć oblicza MAE i MSE.

Odpowiedzi:

1. B
2. B
3. 0,40
4. C, d
5. C

6. 66, 75, 50
7. 7
8. A, b
9. C
10. B
11. B, c
12. Kolejno:
    a. 0,18
    b. 0,03
    c. 0,17
    d. 0,28
    e. 0,15
13. A, b
14. D
15. B
16. E
17. A, d
18. B
19. B
20. A, B, C
21. A, B, D
22. 5
23. 5,6, nie