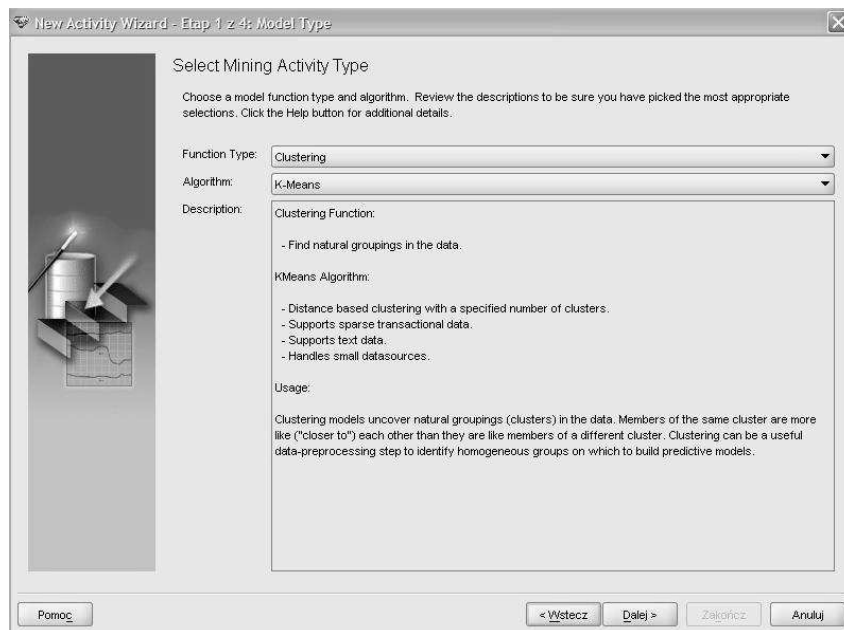


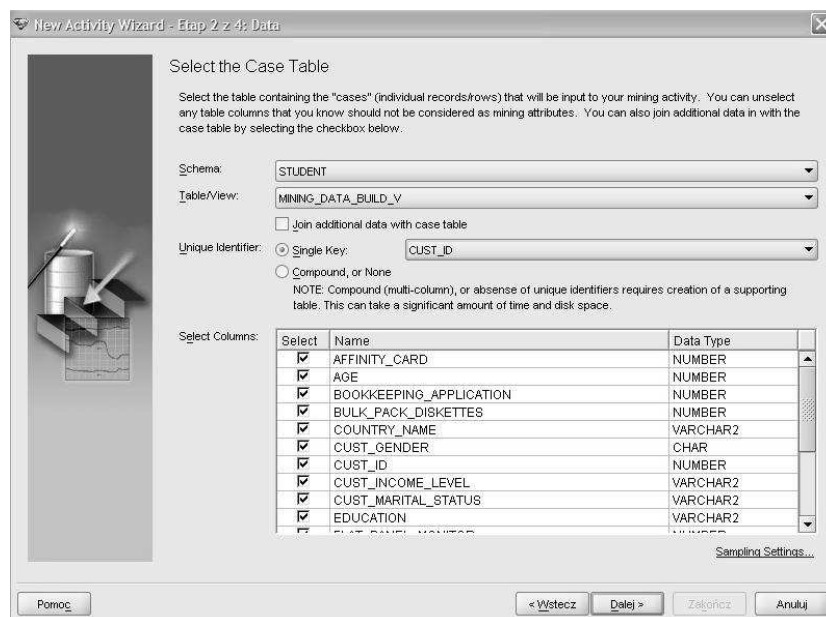
Laboratorium 8

Grupowanie z wykorzystaniem algorytmu K-Means.

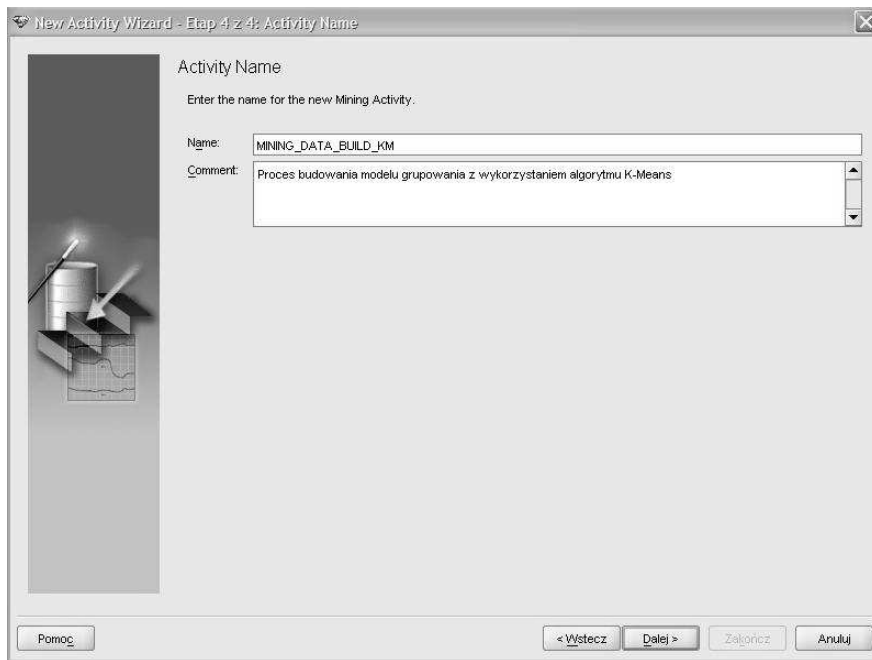
1. Uruchom narzędzie Oracle Data Miner i połącz się z serwerem bazy danych.
2. Z menu głównego wybierz Activity→Build. Na ekranie powitalnym kliknij przycisk Dalej>.
3. Z listy Function Type wybierz Clustering. Rozwiń listę Algorithm i wybierz z niej algorytm K-Means. Kliknij przycisk Dalej>.



4. Wskaż schemat STUDENT i tabelę MINING_DATA_BUILD_V jako źródło danych do eksploracji. Jako klucz podstawowy wskaż atrybut CUST_ID. Kliknij przycisk Dalej>.



5. Podaj nazwę i krótki opis procesu eksploracji. Kliknij przycisk **Dalej**.



New Activity Wizard - Etap 4 z 4: Activity Name

Activity Name

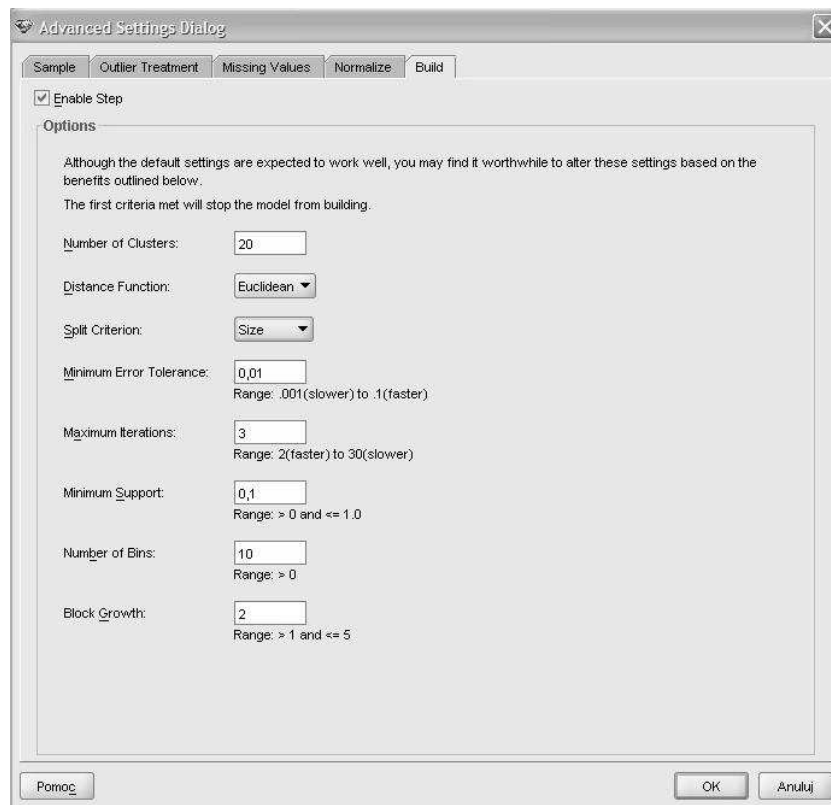
Enter the name for the new Mining Activity.

Name: MINING_DATA_BUILD_KM

Comment: Proces budowania modelu grupowania z wykorzystaniem algorytmu K-Means

Pomoc < Wstecz Dalej > Zakończ Anuluj

6. Kliknij przycisk **Advanced Settings**. Upewnij się, że na zakładce **Sample** opcja próbkowania jest wyłączona (pole wyboru **Enable Step** jest odznaczone). Analogicznie upewnij się, że wyłączone są kroki **Outlier Treatment**, **Missing Values** i **Normalize**. Przejdź na zakładkę **Build**, wpisz wartość 20 jako liczba klastrów (**Number of Clusters**). Wybierz euklidesową funkcję odległości (**Distance Function - Euclidean**), oraz **Split Criterion** ustaw na **Size**. Kliknij przycisk **OK**.



Advanced Settings Dialog

Sample Outlier Treatment Missing Values Normalize Build

Enable Step

Options

Although the default settings are expected to work well, you may find it worthwhile to alter these settings based on the benefits outlined below.

The first criteria met will stop the model from building.

Number of Clusters: 20

Distance Function: Euclidean

Split Criterion: Size

Minimum Error Tolerance: 0,01
Range: .001(slower) to .1(faster)

Maximum Iterations: 3
Range: 2(faster) to 30(slower)

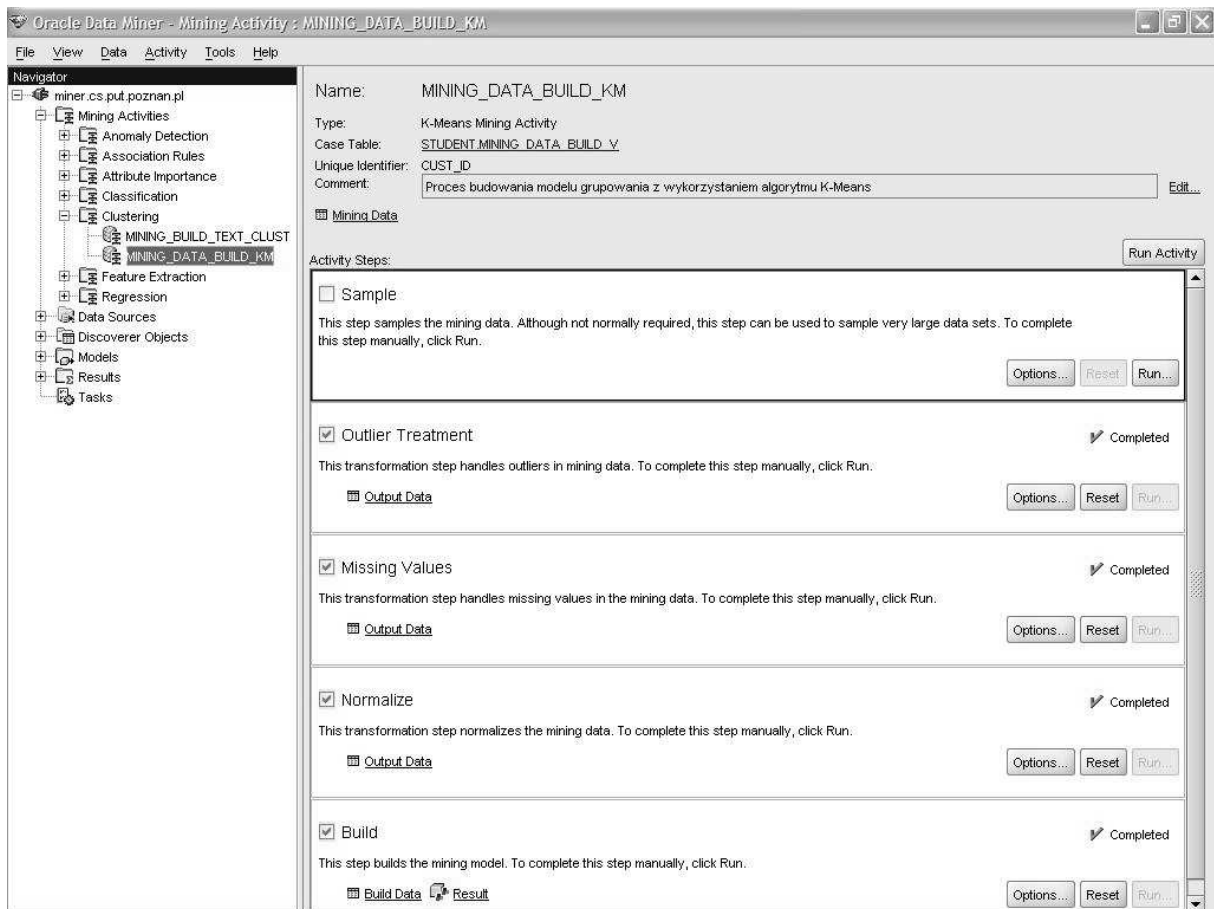
Minimum Support: 0,1
Range: > 0 and <= 1.0

Number of Bins: 10
Range: > 0

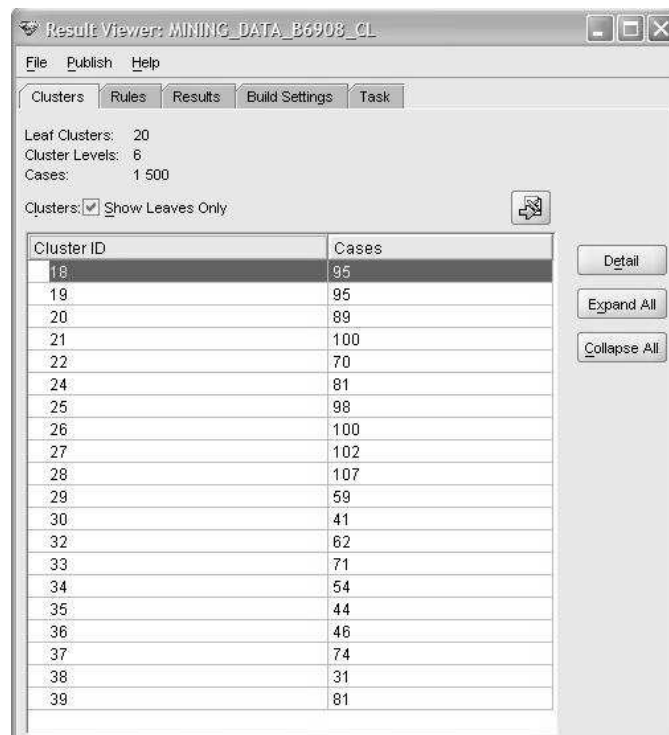
Block Growth: 2
Range: > 1 and <= 5

Pomoc OK Anuluj

7. Upewnij się, że opcja Run upon finish jest włączona. Kliknij przycisk **Zakończ**.



8. Kliknij na odnośnik Result w bloku Build. Zaznacz opcję Show Leaves Only.



9. Przejdź na zakładkę Rules. Zaznacz opcję Only Show Rules for Leaf Clusters. Wybierz dowolny klastery i przeanalizuj atrybuty, które trafiają do wybranego klastra.

Result Viewer: MINING_DATA_B6908_CL

File Publish Help

Clusters Rules Results Build Settings Task

Show topmost relevant attributes: 10

Refresh

Rules Only Show Rules for Leaf Clusters

Cluster ID	Confidence (%)	Support Count
18	82,1052631579	78
19	82,1052631579	78
20	85,393258427	76
21	75,0000	75
22	82,8571428571	58
24	86,4197530864	70
25	81,6326530612	80
26	74,0000	74
27	75,4901960784	77
28	76,6355140187	82
29	83,0508474576	49
30	85,3658536585	35
32	82,2580645161	51
33	77,4647887324	55
34	85,1851851852	46
35	75,0000	33
36	78,2608695652	36
37	86,4864864865	64
38	80,6451612903	25
39	82,7160493827	67

Rule Detail

IF
 AFFINITY_CARD in (0.0) and AGE <= 67.4 and AGE >= 23.3 and BOOKKEEPING_APPLICATION in (1.0) and BULK_PACK_DISKETTES in (1.0) and COUNTRY_NAME in (United States of America) and CUST_GENDER in (M) and CUST_INCOME_LEVEL in (I: 170,000 - 189,999, J: 190,000 - 249,999, K: 250,000 - 299,999, L: 300,000 and above) and CUST_MARITAL_STATUS in (Married) and EDUCATION in (HS-grad) and FLAT_PANEL_MONITOR in (1.0) and HOME_THEATER_PACKAGE in (0.0,1.0) and HOUSEHOLD_SIZE in (3.0) and OCCUPATION in (?, Cleric., Crafts, Exec., Handler, Machine, Other, Sales, Transp.) and OS_DOC_SET_KANJI in (0.0) and YRS_RESIDENCE <= 7.2 and YRS_RESIDENCE >= 2.6999999999999997 and Y_BOX_GAMES in (0.0)

THEN
 Cluster equal 18

Confidence (%)=82.1052631578947
 Support =78

Ćwiczenie samodzielne

Na podstawie tabeli PRACOWNICY zbuduj następującą perspektywę:

```
SQL> desc pracownicy_v
Name                               Null?      Type
-----
NAZWISKO                           VARCHA2(15)
ETAT                                 VARCHA2(10)
ROK                                  NUMBER
PLACA_POD                            NUMBER(6,2)
PLACA_DOD                            NUMBER(6,2)
ID_ZESP                              NUMBER(2)
ID_SZEFA                             NUMBER(4)
ID_Prac                             NOT NULL  NUMBER(4)
```

Utworzoną przez siebie perspektywę wykorzystaj do zbudowania modelu znajdującego n grup najbardziej podobnych do siebie pracowników (wartość n wyznacz eksperymentalnie). Do znalezienia grup zastosuj algorytm k-Means. Wyświetl informacje o uzyskanym modelu.

Aby zobaczyć działanie algorytmu, zastosuj zbudowany przez siebie model do zawartości perspektywy PRACOWNICY_V. Zwróć uwagę, jak algorytm przydziela profesorów do grup. Ostatecznym celem ćwiczenia jest napisanie zapytania, które wyświetli przypisanie

pracowników do poszczególnych grup (wraz z prawdopodobieństwem przypisania pracownika do danej grupy). Postaraj się uzyskać podobny efekt:

NAZWISKO	ETAT	CLUSTER_ID	PROBABILITY
BLAZEWICZ	PROFESOR	4	.9972
SLOWINSKI	PROFESOR	4	.904
WEGLARZ	DYREKTOR	6	.9999
BRZEZINSKI	PROFESOR	7	.9343
MORZY	PROFESOR	7	.9254
KROLIKOWSKI	ADIUNKT	8	.6857
BIALY	STAZYSTA	8	.9286
ZAKRZEWICZ	STAZYSTA	8	.9791
KOSZLAJDA	ADIUNKT	8	.8453
JEZIERSKI	ASYSTENT	8	.7038
MATYSIAK	ASYSTENT	8	.904
MAREK	SEKRETARKA	9	.9575
KONOPKA	ASYSTENT	9	.9036
HAPKE	ASYSTENT	9	.598

Uwaga:

Pamiętaj, aby dane przetwarzane przez algorytm k-Means poddać normalizacji!