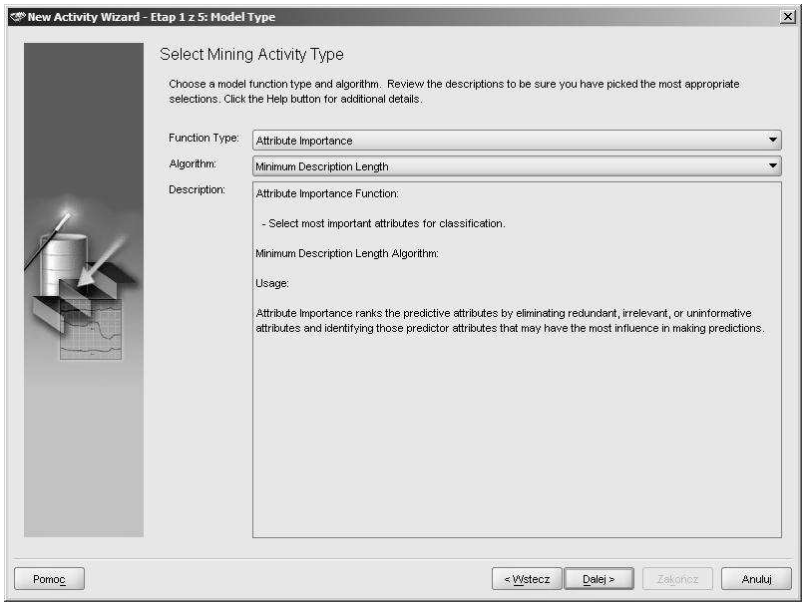
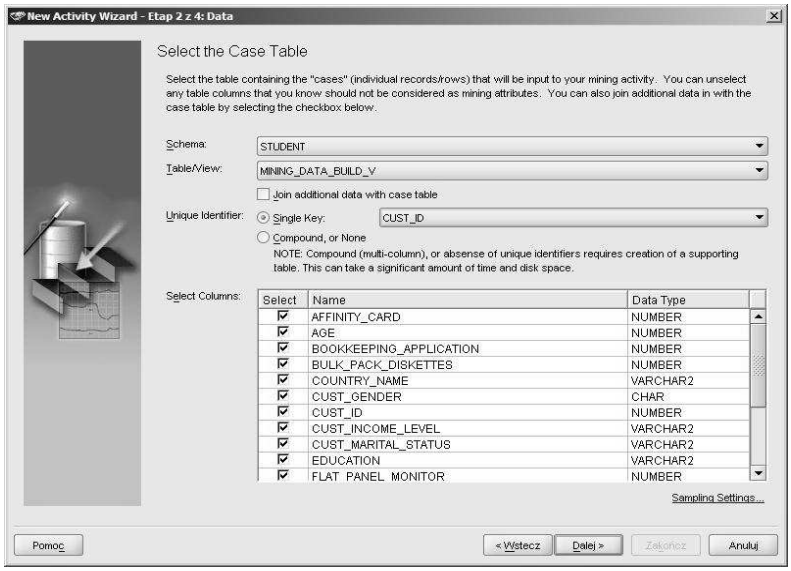
# Laboratorium 2

## *Określanie ważności atrybutów.*

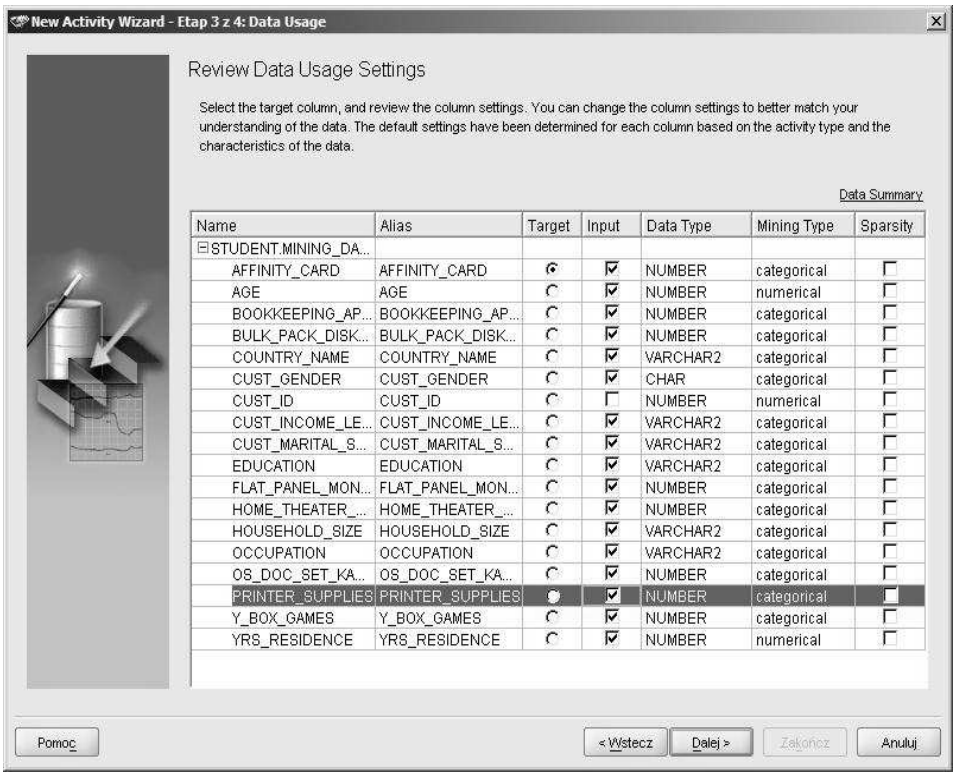
1. Uruchom narzędzie Oracle Data Miner i połącz się z serwerem bazy danych.
2. Z menu głównego wybierz Activity→Build. Na ekranie powitalnym kliknij przycisk Dalej>.
3. Z listy Function Type wybierz Attribute Importance. Jedyny dostępny algorytm, Minimum Description Length, zostaje automatycznie wybrany jako algorytm przetwarzający. Kliknij przycisk Dalej>.



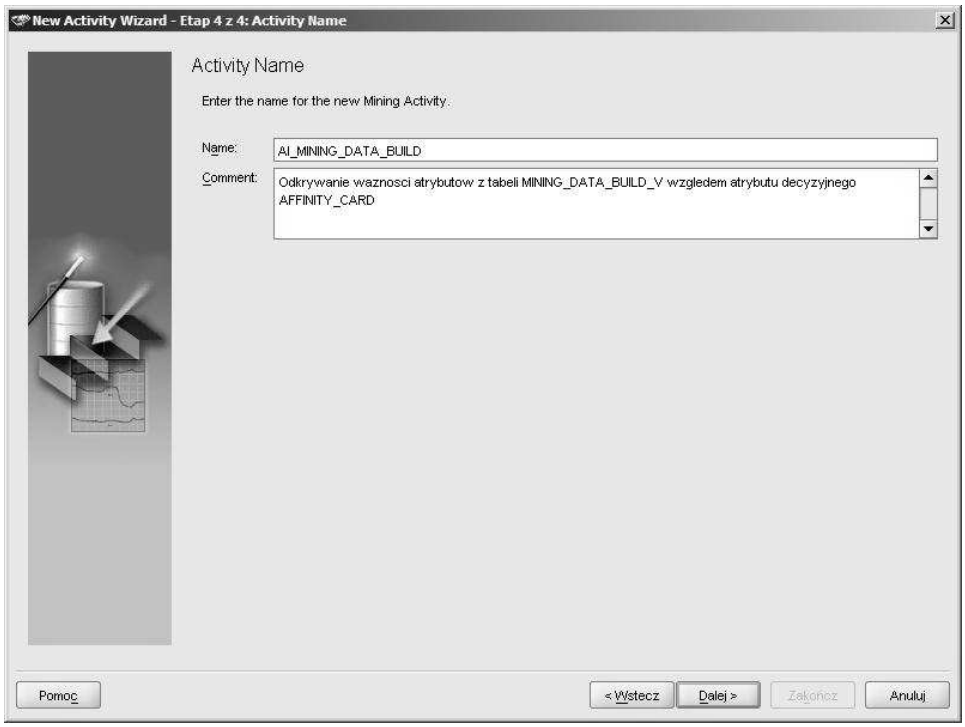

4. Z listy dostępnych schematów wybierz schemat STUDENT. Wybierz tabelę MINING_DATA_BUILD_V. Jako klucz podstawowy (Unique identifier) zaznacz przycisk radiowy Single Key i z listy wybierz atrybut CUST_ID. Upewnij się, że wszystkie atrybuty na liście Selected Columns są wybrane. Kliknij przycisk Dalej>.



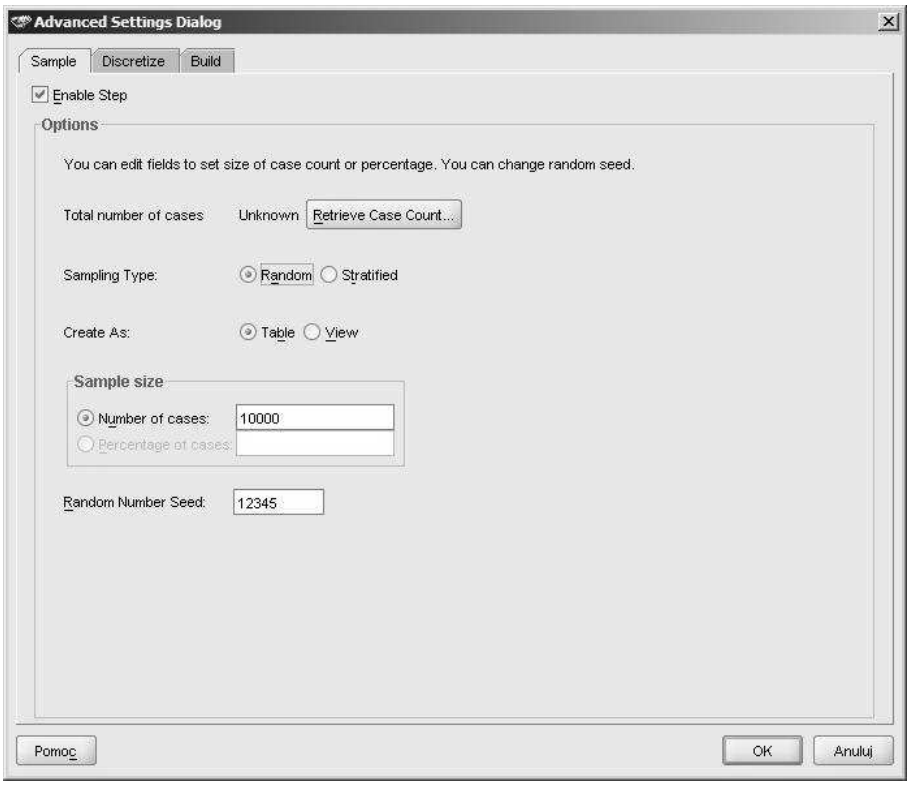
5. Kolumna Target zawiera informację o atrybucie decyzyjnym, względem którego staramy się określić ważność pozostałych atrybutów (predyktorów). Wskaż atrybut `AFFINITY_CARD` (oznacza, czy dany klient skorzysta z karty lojalnościowej) jako atrybut decyzyjny w kolumnie Target. Atrybut `PRINTER_SUPPLIES` został automatycznie odrzucony, ponieważ zawiera tylko jedną wartość. Włącz atrybut `PRINTER_SUPPLIES` do przetwarzania. Kliknij przycisk Dalej>.



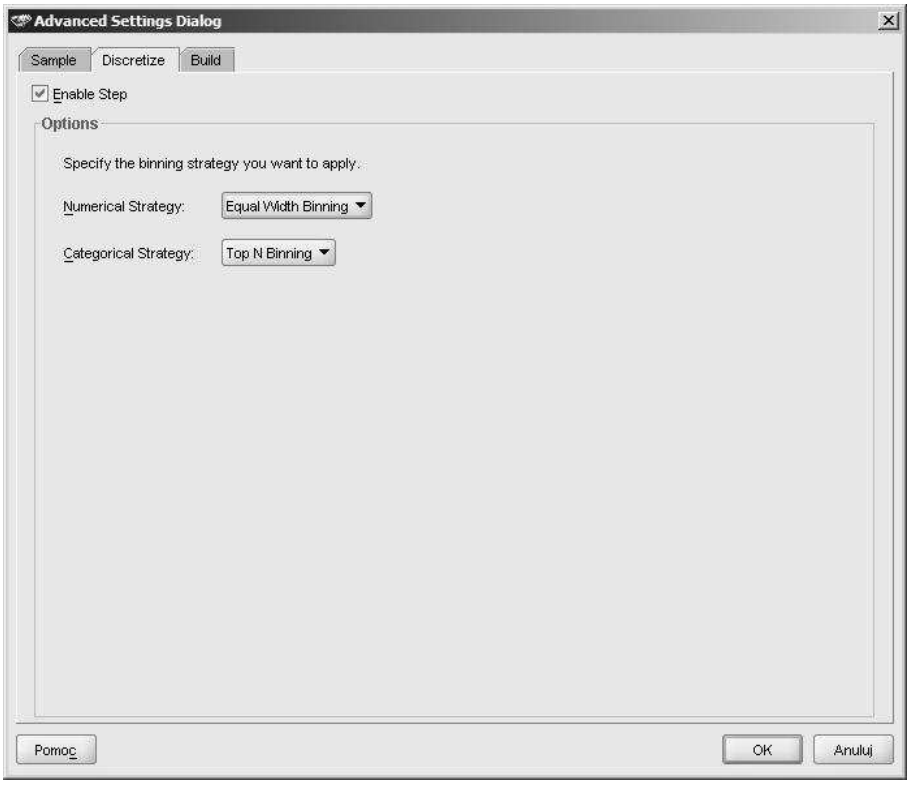

6. Wprowadź nazwę i opis dla procesu eksploracji. Kliknij przycisk Dalej>.



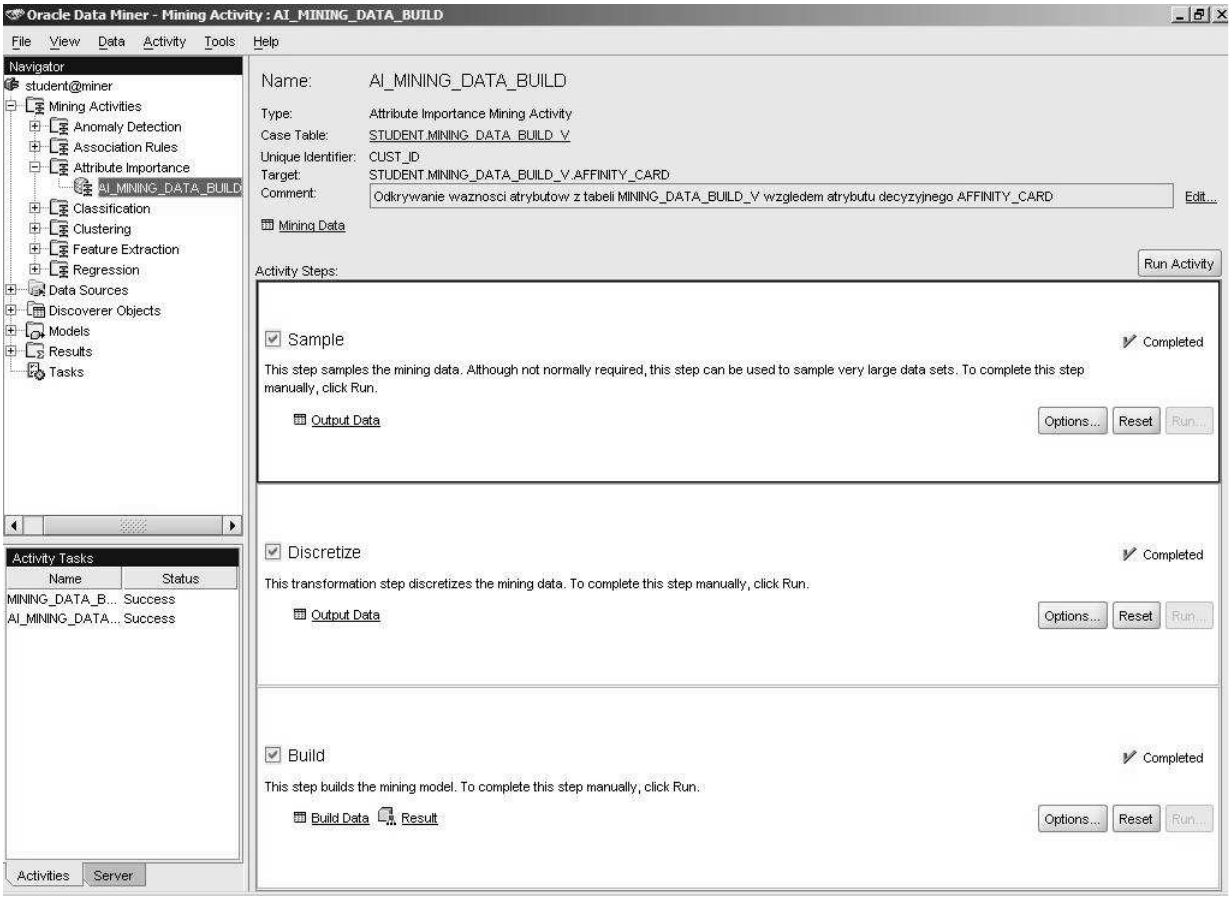
7. Kliknij przycisk Advanced Settings. Na zakładce Sample włącz opcję Enable Step. Zmień rodzaj próbkowania na losowy (opcja Sampling Type, wartość Random).



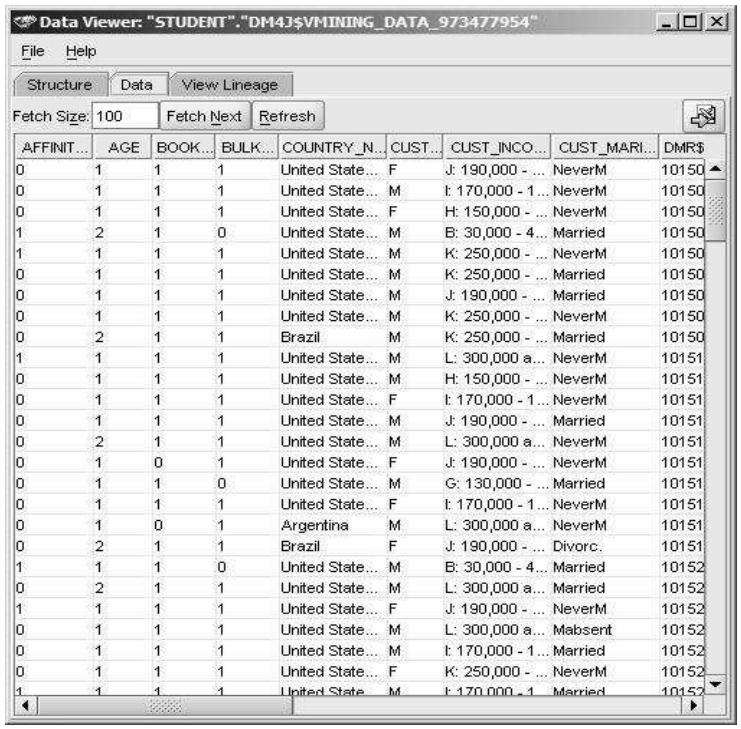

8. Przejdź na zakładkę Discretize. Upewnij się, że dyskretyzacja jest włączona i zostanie wykonana automatycznie przed uruchomieniem się właściwego algorytmu. Zmień sposób dyskretyzacji atrybutów numerycznych na Equal Width Binning.



9. Kliknij przycisk OK. Upewnij się, że jest zaznaczona opcja Run upon finish. Kliknij klawisz Zakończ. Ekran powinien mieć w tej chwili następujący wygląd. Bloki w prawym panelu reprezentują następujące po sobie etapy procesu eksploracji danych. Dla każdego etapu można zobaczyć dane stanowiące wynik działania danego etapu (odnośnik Output Data). Przycisk Run Activity powoduje uruchomienie całego procesu eksploracji. Istnieje też możliwość zmiany parametrów i usunięcia poprzednich wyników dla dowolnego procesu i uruchomienie całej eksploracji w sposób przyrostowy (przyciski Reset, Options i Run).

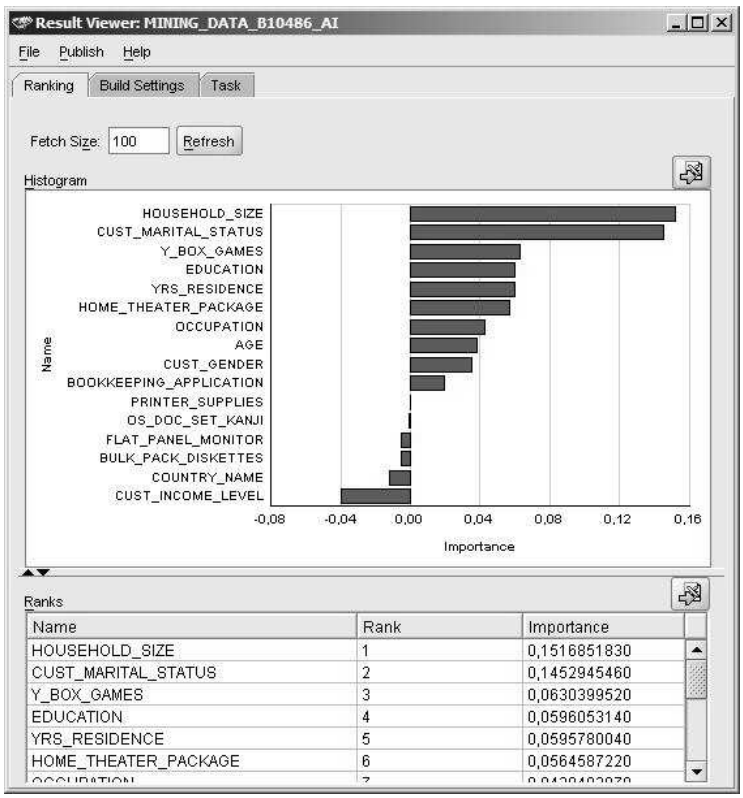



10. Kliknij na odnośnik Output Data w bloku Discretize. Kliknij na odnośnik do tymczasowej tabeli. W nowym okienku kliknij na zakładce Data. Zwróć uwagę, że atrybuty numeryczne (`AGE`, `YRS_RESIDENCE`) zostały automatycznie dyskretyzowane. Zamknij okienko z tabelą tymczasową.




| AFFINIT... | AGE | BOOK... | BULK... | COUNTRY_N... | CUST... | CUST_INCO... | CUST_MARI... | DMR$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | United State... | F | J: 190,000 - ... | NeverM | 10150 |
| 0 | 1 | 1 | 1 | United State... | M | I: 170,000 - 1... | NeverM | 10150 |
| 0 | 1 | 1 | 1 | United State... | F | H: 150,000 - ... | NeverM | 10150 |
| 1 | 2 | 1 | 0 | United State... | M | B: 30,000 - 4... | Married | 10150 |
| 1 | 1 | 1 | 1 | United State... | M | K: 250,000 - ... | NeverM | 10150 |
| 0 | 1 | 1 | 1 | United State... | M | K: 250,000 - ... | Married | 10150 |
| 0 | 1 | 1 | 1 | United State... | M | J: 190,000 - ... | Married | 10150 |
| 0 | 1 | 1 | 1 | United State... | M | K: 250,000 - ... | NeverM | 10150 |
| 0 | 2 | 1 | 1 | Brazil | M | K: 250,000 - ... | Married | 10150 |
| 1 | 1 | 1 | 1 | United State... | M | L: 300,000 a... | NeverM | 10151 |
| 0 | 1 | 1 | 1 | United State... | M | H: 150,000 - ... | NeverM | 10151 |
| 0 | 1 | 1 | 1 | United State... | F | I: 170,000 - 1... | NeverM | 10151 |
| 0 | 1 | 1 | 1 | United State... | M | J: 190,000 - ... | Married | 10151 |
| 0 | 2 | 1 | 1 | United State... | M | L: 300,000 a... | NeverM | 10151 |
| 0 | 1 | 0 | 1 | United State... | F | J: 190,000 - ... | NeverM | 10151 |
| 0 | 1 | 1 | 0 | United State... | M | G: 130,000 - ... | Married | 10151 |
| 0 | 1 | 1 | 1 | United State... | F | I: 170,000 - 1... | NeverM | 10151 |
| 0 | 1 | 0 | 1 | Argentina | M | L: 300,000 a... | NeverM | 10151 |
| 0 | 2 | 1 | 1 | Brazil | F | J: 190,000 - ... | Divorc. | 10151 |
| 1 | 1 | 1 | 0 | United State... | M | B: 30,000 - 4... | Married | 10152 |
| 0 | 2 | 1 | 1 | United State... | M | L: 300,000 a... | Married | 10152 |
| 1 | 1 | 1 | 1 | United State... | F | J: 190,000 - ... | NeverM | 10152 |
| 0 | 1 | 1 | 1 | United State... | M | L: 300,000 a... | Mabsent | 10152 |
| 0 | 1 | 1 | 1 | United State... | M | I: 170,000 - 1... | Married | 10152 |
| 0 | 1 | 1 | 1 | United State... | F | K: 250,000 - ... | NeverM | 10152 |
| 1 | 1 | 1 | 1 | United State... | M | I: 170,000 - 1... | Married | 10152 |

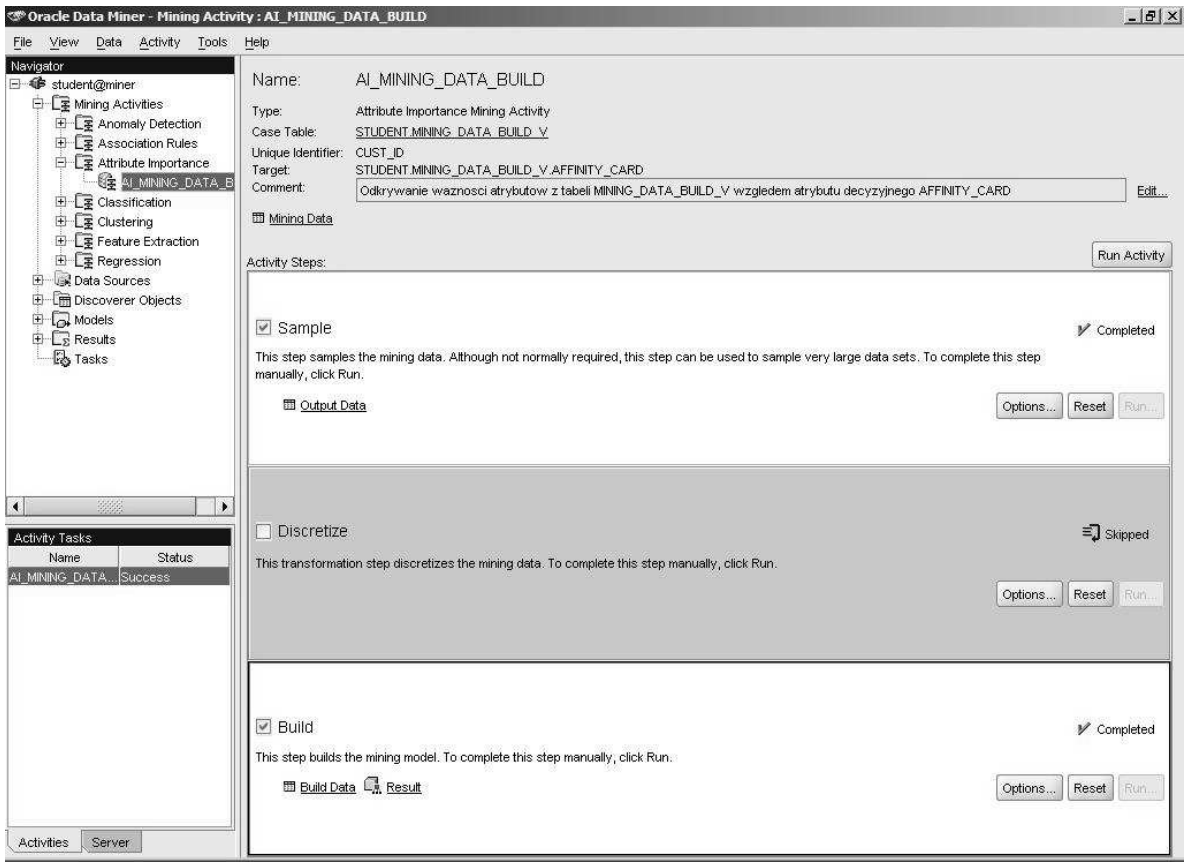
11. Kliknij na odnośnik Result w bloku Build. Przeanalizuj uzyskane wyniki. Które atrybuty są najbardziej przydatne do predykcji wartości atrybutu decyzyjnego? Atrybuty o ważności mniejszej niż 0 są uznawane za całkowicie nieprzydatne. Jaką ważność ma atrybut PRINTER_SUPPLIES posiadający tylko jedną wartość?




Ranks

| Name | Rank | Importance |
|---|---|---|
| HOUSEHOLD_SIZE | 1 | 0,1516851830 |
| CUST_MARITAL_STATUS | 2 | 0,1452945460 |
| Y_BOX_GAMES | 3 | 0,0630399520 |
| EDUCATION | 4 | 0,0596053140 |
| YRS_RESIDENCE | 5 | 0,0595780040 |
| HOME_THEATER_PACKAGE | 6 | 0,0564587220 |

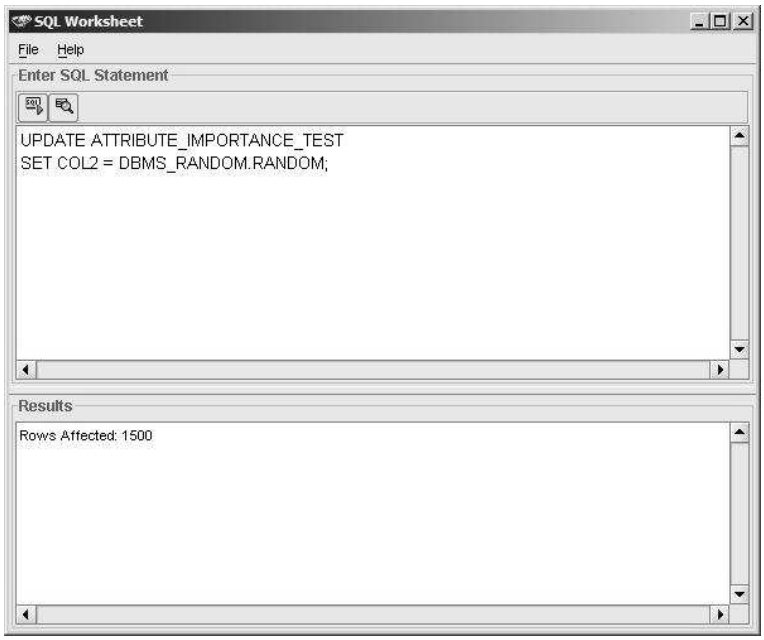
12. Kliknij na  umieszczoną w prawym górnym rogu. Jako format pliku graficznego do eksportu wybierz GIF. Kliknij przycisk OK. Zapisz plik na dysku lokalnym. Obejrzyj wygenerowany plik. Zamknij okienko prezentujące wynik działania algorytmu.
13. W bloku Sample kliknij przycisk Reset. W okienku z pytaniem o potwierdzenie usunięcia wyniku działania kroku próbkowania (i wszystkich kroków następujących po próbkowaniu) kliknij przycisk Tak. Wyniki wszystkich kroków zostaną usunięte. W bloku Sample kliknij przycisk Options i zmień sposób próbkowania na próbkowanie warstwowane (opcja Sampling Type, wartość Stratified). Kliknij przycisk OK. Odznacz pole wyboru obok nazwy bloku Discretize (w ten sposób w trakcie kolejnego uruchomienia algorytmu nie zostanie przeprowadzona automatyczna dyskretyzacja). Kliknij przycisk Run Activity umieszczony w prawej górnej części ekranu. Po wykonaniu tych czynności ekran powinien wyglądać jak na obrazku poniżej. Porównaj uzyskany wynik z poprzednim wynikiem. Czy obserwujesz jakieś zmiany?



14. Z menu głównego wybierz Tools→SQL Worksheet. W górnym oknie wprowadź po kolei następujące komendy:

```
CREATE TABLE ATTRIBUTE_IMPORTANCE_TEST AS
SELECT * FROM MINING_DATA_BUILD_V;

ALTER TABLE ATTRIBUTE_IMPORTANCE_TEST
ADD COL1 NUMBER;

ALTER TABLE ATTRIBUTE_IMPORTANCE_TEST
ADD COL2 NUMBER;

UPDATE ATTRIBUTE_IMPORTANCE_TEST
SET COL1 = 1 - AFFINITY_CARD;

UPDATE ATTRIBUTE_IMPORTANCE_TEST
SET COL2 = DBMS_RANDOM.RANDOM;

COMMIT;
```

15. W nowo utworzonej tabeli znajdują się dwa dodatkowe atrybuty: `COL1` i `COL2`. Atrybut `COL1` jest liniowo zależny od atrybutu decyzyjnego (czyli pozwala ze 100% dokładnością przewidzieć wartości atrybutu decyzyjnego), podczas gdy atrybut `COL2` jest całkowicie losowy (czyli nie posiada żadnej wartości informacyjnej).




16. Utwórz nowe zadanie eksploracji (z głównego menu wybierz Activity→Build) i określ ważność atrybutów względem atrybutu decyzyjnego `AFFINITY_CARD` w tabeli `ATTRIBUTE_IMPORTANCE_TEST`. W procesie eksploracji wykorzystaj próbkowanie i automatyczną dyskretyzację atrybutów. Sprawdź, w którym miejscu w rankingu atrybutów znajdą się atrybuty `COL1` i `COL2`.

17. Połącz się z bazą danych wykorzystując iSQLPlus. Wykonaj skrypt `ai.sql`. Po każdym kroku przeanalizuj uzyskane wyniki (komentarz jest umieszczony wewnątrz skryptu).

## Ćwiczenie samodzielne

W schemacie użytkownika **sh** znajdują się tabele `CUSTOMERS`,`COUNTRIES`,`SALES` o następujących schematach:

```
SQL> desc sh.customers
 Name                                      Null?    Type
 ----------------------------------------- -------- ------------
 CUST_ID                                   NOT NULL NUMBER
 CUST_FIRST_NAME                           NOT NULL VARCHAR2(20)
 CUST_LAST_NAME                            NOT NULL VARCHAR2(40)
 CUST_GENDER                               NOT NULL CHAR(1)
 CUST_YEAR_OF_BIRTH                        NOT NULL NUMBER(4)
 CUST_MARITAL_STATUS                                VARCHAR2(20)
 CUST_STREET_ADDRESS                       NOT NULL VARCHAR2(40)
 CUST_POSTAL_CODE                          NOT NULL VARCHAR2(10)
 CUST_CITY                                 NOT NULL VARCHAR2(30)
 CUST_CITY_ID                              NOT NULL NUMBER
 CUST_STATE_PROVINCE                       NOT NULL VARCHAR2(40)
 CUST_STATE_PROVINCE_ID                    NOT NULL NUMBER
 COUNTRY_ID                                NOT NULL NUMBER
 CUST_MAIN_PHONE_NUMBER                    NOT NULL VARCHAR2(25)
 CUST_INCOME_LEVEL                                  VARCHAR2(30)
 CUST_CREDIT_LIMIT                                  NUMBER
 CUST_EMAIL                                         VARCHAR2(30)
 CUST_TOTAL                                NOT NULL VARCHAR2(14)
 CUST_TOTAL_ID                             NOT NULL NUMBER
 CUST_SRC_ID                                        NUMBER
 CUST_EFF_FROM                                      DATE
 CUST_EFF_TO                                        DATE
 CUST_VALID                                         VARCHAR2(1)

 SQL> desc sh.countries
 Name                                      Null?    Type
 ----------------------------------------- -------- -------------
 COUNTRY_ID                                NOT NULL NUMBER
 COUNTRY_ISO_CODE                          NOT NULL CHAR(2)
 COUNTRY_NAME                              NOT NULL VARCHAR2(40)
 COUNTRY_SUBREGION                         NOT NULL VARCHAR2(30)
 COUNTRY_SUBREGION_ID                      NOT NULL NUMBER
 COUNTRY_REGION                            NOT NULL VARCHAR2(20)
 COUNTRY_REGION_ID                         NOT NULL NUMBER
 COUNTRY_TOTAL                             NOT NULL VARCHAR2(11)
 COUNTRY_TOTAL_ID                          NOT NULL NUMBER
 COUNTRY_NAME_HIST                                  VARCHAR2(40)

SQL> desc sh.sales
 Name                                      Null?    Type
 ----------------------------------------- -------- ------------
 PROD_ID                                   NOT NULL NUMBER
 CUST_ID                                   NOT NULL NUMBER
 TIME_ID                                   NOT NULL DATE
 CHANNEL_ID                                NOT NULL NUMBER
 PROMO_ID                                  NOT NULL NUMBER
 QUANTITY_SOLD                             NOT NULL NUMBER(10,2)
 AMOUNT_SOLD                               NOT NULL NUMBER(10,2)
```

Stwórz we własnym schemacie perspektywę przedstawiającą wszystkie informacje o klientach. Perspektywa powinna zawierać wszystkie atrybuty opisujące klienta (wraz z atrybutami opisującymi miasto i kraj klienta). Dodatkowo, perspektywa powinna zawierać atrybut `FLAG`, którego wartością jest 1 (jeżeli dany klient dokonał więcej niż 200 zakupów)

lub 0 (jeżeli dany klient dokonał mniej niż 200 zakupów). Oczywiście liczbę zakupów można wyznaczyć na podstawie zawartości tabeli SALES. Zastanów się, które atrybuty, Twoim zdaniem, najlepiej nadają się do przewidywania czy dany klient będzie kupował dużo towarów, a które nie niosą takiej informacji wcale. Następnie wykorzystaj Oracle Data Mining PL/SQL API lub narzędzie Oracle Data Miner do określenia ważności wszystkich atrybutów względem atrybutu docelowego FLAG. Na ile Twoja intuicja się sprawdziła?

Na potrzeby eksploracji dane wejściowe powinny znajdować się w schemacie:

```
SQL> desc customers_v
 Name                                      Null?    Type
 ----------------------------------------- -------- ------------
 COUNTRY_ID                                NOT NULL NUMBER
 CUST_ID                                   NOT NULL NUMBER
 CUST_FIRST_NAME                           NOT NULL VARCHAR2(20)
 CUST_LAST_NAME                            NOT NULL VARCHAR2(40)
 CUST_GENDER                               NOT NULL CHAR(1)
 CUST_YEAR_OF_BIRTH                        NOT NULL NUMBER(4)
 CUST_MARITAL_STATUS                                VARCHAR2(20)
 CUST_STREET_ADDRESS                       NOT NULL VARCHAR2(40)
 CUST_POSTAL_CODE                          NOT NULL VARCHAR2(10)
 CUST_CITY                                 NOT NULL VARCHAR2(30)
 CUST_CITY_ID                              NOT NULL NUMBER
 CUST_STATE_PROVINCE                       NOT NULL VARCHAR2(40)
 CUST_STATE_PROVINCE_ID                    NOT NULL NUMBER
 CUST_MAIN_PHONE_NUMBER                    NOT NULL VARCHAR2(25)
 CUST_INCOME_LEVEL                                  VARCHAR2(30)
 CUST_CREDIT_LIMIT                                  NUMBER
 CUST_EMAIL                                         VARCHAR2(30)
 CUST_TOTAL                                NOT NULL VARCHAR2(14)
 CUST_TOTAL_ID                             NOT NULL NUMBER
 CUST_SRC_ID                                        NUMBER
 CUST_VALID                                         VARCHAR2(1)
 COUNTRY_ISO_CODE                          NOT NULL CHAR(2)
 COUNTRY_NAME                              NOT NULL VARCHAR2(40)
 COUNTRY_SUBREGION                         NOT NULL VARCHAR2(30)
 COUNTRY_SUBREGION_ID                      NOT NULL NUMBER
 COUNTRY_REGION                            NOT NULL VARCHAR2(20)
 COUNTRY_REGION_ID                         NOT NULL NUMBER
 COUNTRY_TOTAL                             NOT NULL VARCHAR2(11)
 COUNTRY_TOTAL_ID                          NOT NULL NUMBER
 COUNTRY_NAME_HIST                                  VARCHAR2(40)
 FLAG                                               NUMBER
```

**UWAGA:**

- nie kopiuj danych do własnego schematu
- koniecznie pomiń w utworzonej przez siebie perspektywie atrybuty CUST_EFF_FROM i CUST_EFF_FROM, oba atrybuty są typu DATE, który nie jest obsługiwany przez ODM