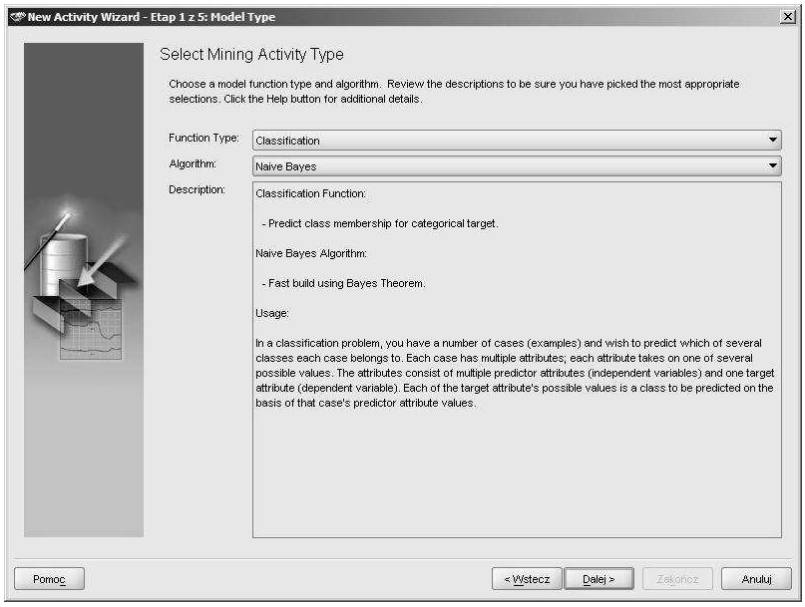
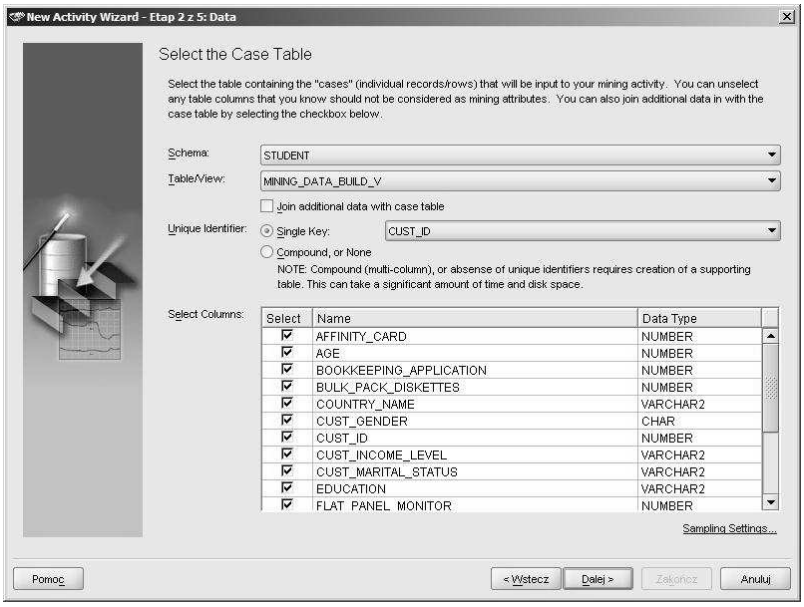
Laboratorium 4

Naiwny klasyfikator Bayesa.

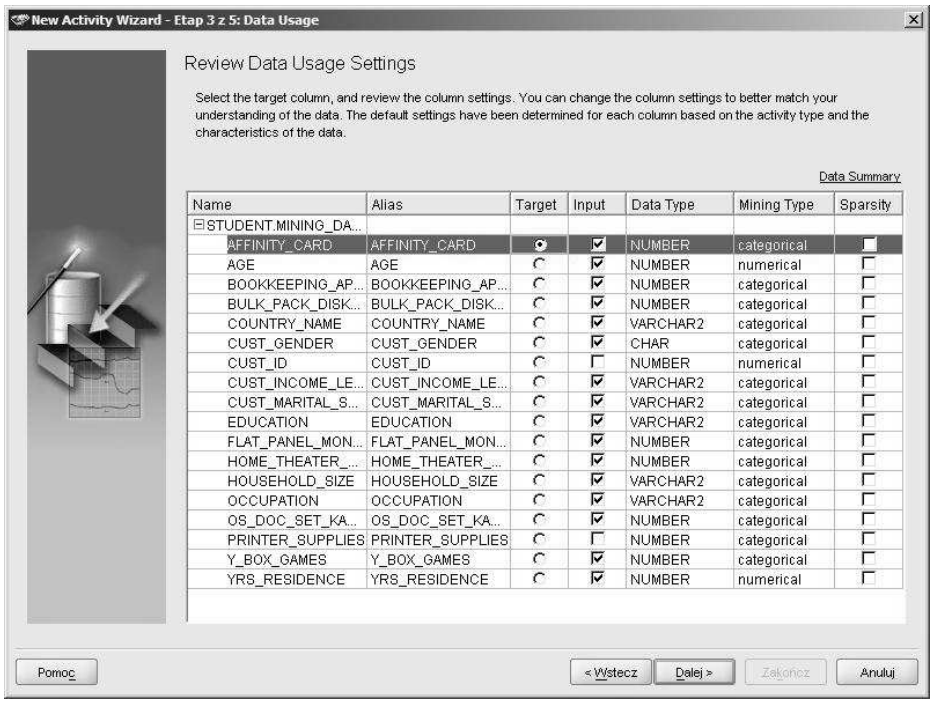
1. Uruchom narzędzie Oracle Data Miner i połącz się z serwerem bazy danych.
2. Z menu głównego wybierz Activity→Build. Na ekranie powitalnym kliknij przycisk Dalej>.
3. Z listy Function Type wybierz Classification. Rozwiń listę Algorithm i wybierz z niej algorytm Naive Bayes. Kliknij przycisk Dalej>.



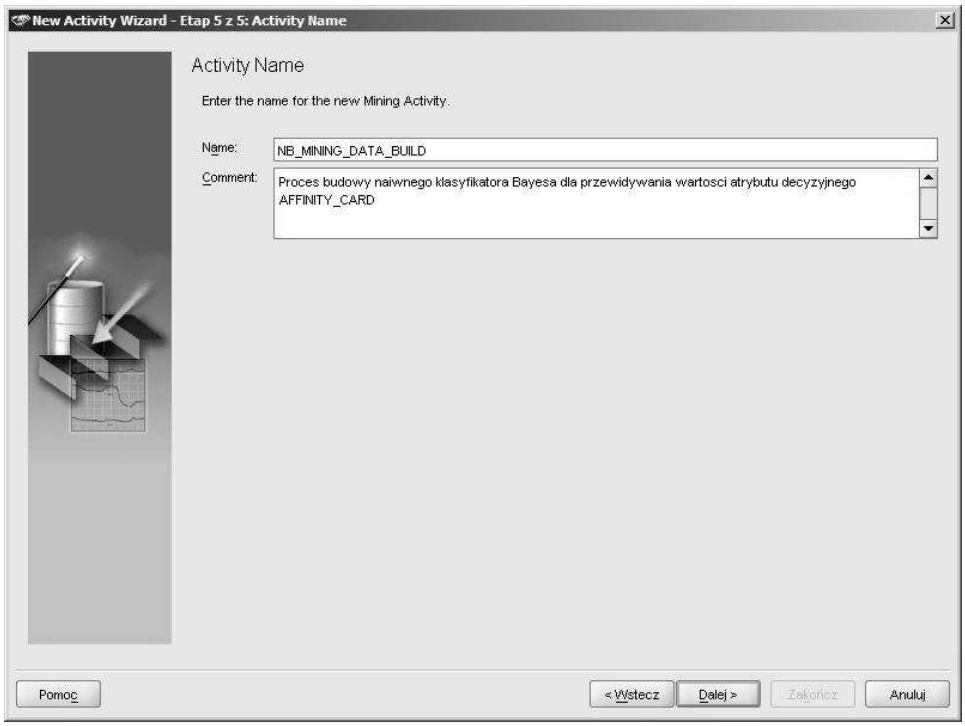
4. Wskaż schemat `STUDENT` i tabelę `MINING_DATA_BUILD_V` jako źródło danych do eksploracji. Jako klucz podstawowy wskaż atrybut `CUST_ID`. Kliknij przycisk Dalej>.



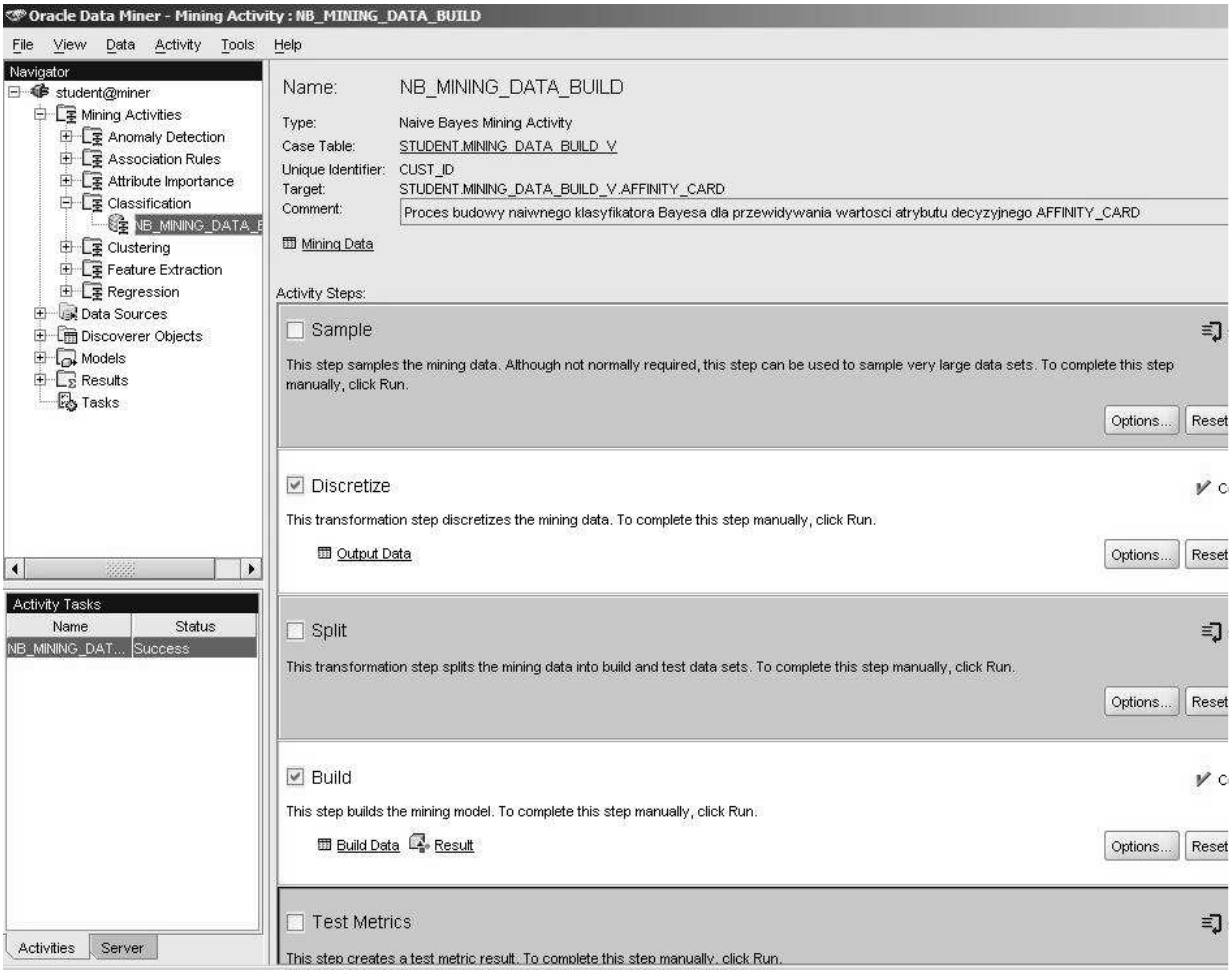
5. Jako atrybut decyzyjny zaznacz atrybut `AFFINITY_CARD` (pole radiowe w kolumnie `Target`). Upewnij się, że atrybuty `CUST_ID` i `PRINTER_SUPPLIES` są wyłączone z eksploracji (są bezwartościowe i nie niosą żadnej informacji). Kliknij przycisk Dalej>.



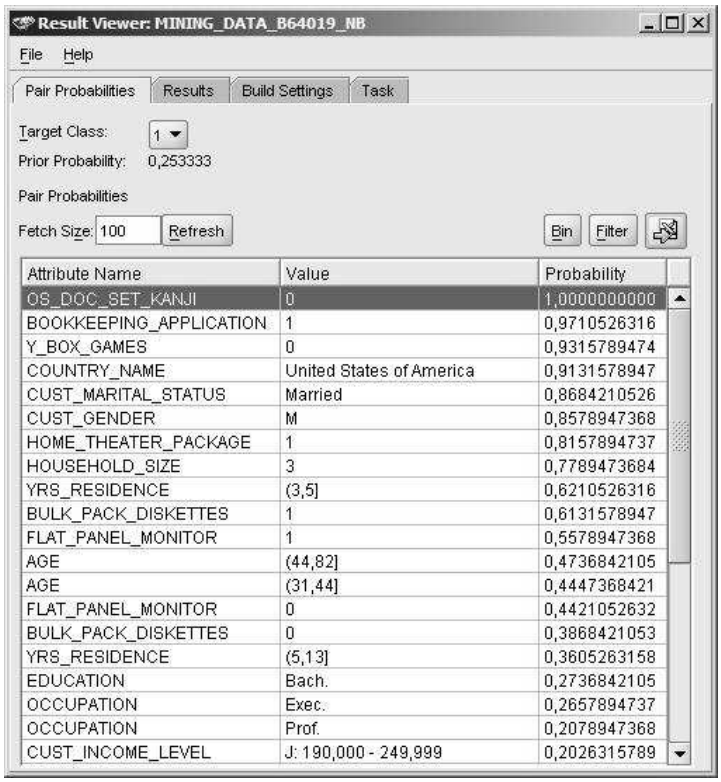
6. Z listy rozwijanej wybierz wartość **1** jako preferowaną wartość atrybutu decyzyjnego (jest to wartość, której poprawne przewidywanie jest najważniejsze, interesuje nas dokładna identyfikacja klientów, którzy prawdopodobnie skorzystają z oferowanej im karty lojalnościowej). Kliknij przycisk Dalej>. Wprowadź nazwę i komentarz do procesu eksploracji. Kliknij przycisk Dalej>.



7. Kliknij przycisk Advanced Settings. Upewnij się, że na zakładce Sample opcja próbkowania jest wyłączona (pole wyboru Enable Step jest odznaczone). Przejdź na zakładkę Discretize. Naiwny klasyfikator Bayesa wymaga, aby atrybuty numeryczne zostały poddane dyskretyzacji. Upewnij się, czy automatyczna dyskretyzacja jest włączona. Przejdź na zakładkę Split i upewnij się, że podział danych wejściowych na zbiór uczący i testujący jest wyłączony. Przejdź na zakładkę Build. Upewnij się, że algorytm będzie się starał osiągnąć maksymalną średnią dokładność (w polu Accuracy Goal wybierz opcję Maximum Average Accuracy). Kliknij na zakładkę Algorithm Settings. Wprowadź wartości parametrów: Singleton Threshold 0.1 i Pairwise Threshold 0.01. Przejdź na zakładkę Test Metrics i wyłącz krok generowania miar oceny dokładności i jakości klasyfikatora. Kliknij przycisk OK. Upewnij się, że opcja Run upon finish jest włączona. Kliknij przycisk Zakończ.

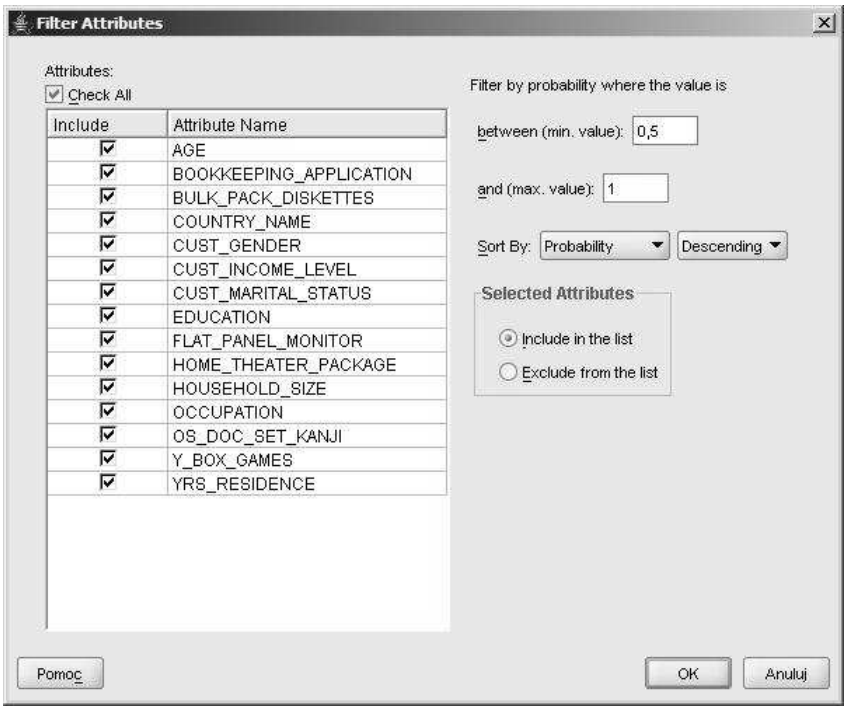


8. Kliknij na odnośnik Results w bloku Build. Na liście rozwijanej Target Class (w lewym górnym rogu okna) wybierz wartość **1**. Przeanalizuj prawdopodobieństwa warunkowe wartości poszczególnych atrybutów względem określonej wartości atrybutu decyzyjnego.

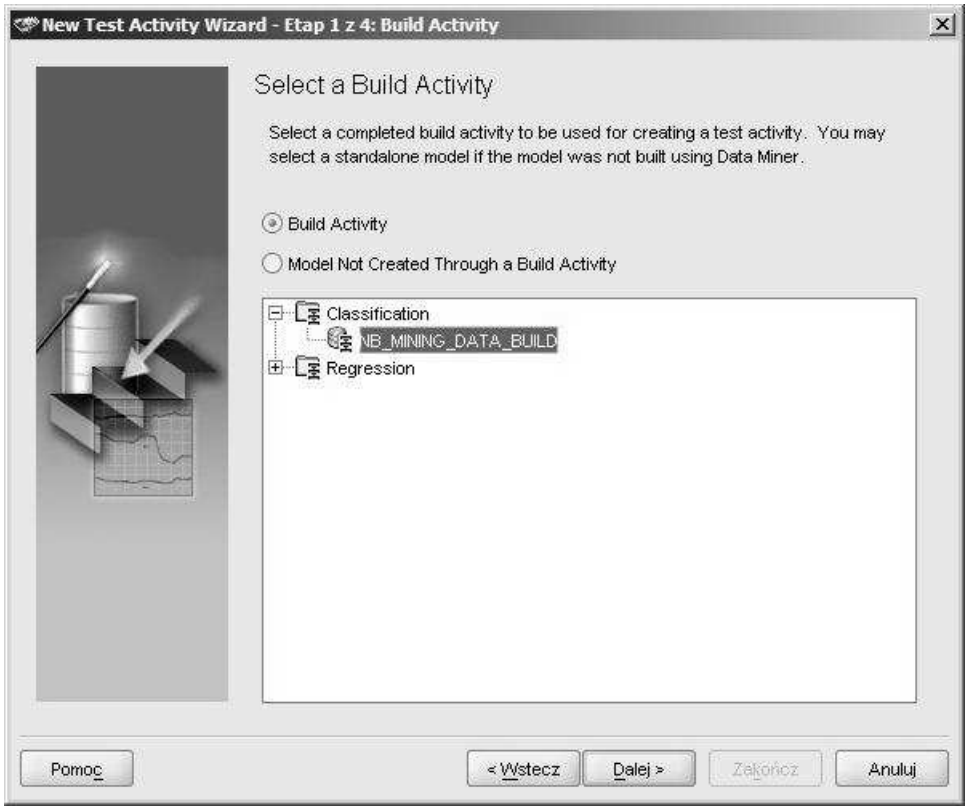


Attribute Name	Value	Probability
OS_DOC_SET_KANJI	0	1,0000000000
BOOKKEEPING_APPLICATION	1	0,9710526316
Y_BOX_GAMES	0	0,9315789474
COUNTRY_NAME	United States of America	0,9131578947
CUST_MARITAL_STATUS	Married	0,8684210526
CUST_GENDER	M	0,8578947368
HOME_THEATER_PACKAGE	1	0,8157894737
HOUSEHOLD_SIZE	3	0,7789473684
YRS_RESIDENCE	(3,5]	0,6210526316
BULK_PACK_DISKETTES	1	0,6131578947
FLAT_PANEL_MONITOR	1	0,5578947368
AGE	(44,82]	0,4736842105
AGE	(31,44]	0,4447368421
FLAT_PANEL_MONITOR	0	0,4421052632
BULK_PACK_DISKETTES	0	0,3868421053
YRS_RESIDENCE	(5,13]	0,3605263158
EDUCATION	Bach.	0,2736842105
OCCUPATION	Exec.	0,2657894737
OCCUPATION	Prof.	0,2078947368
CUST_INCOME_LEVEL	J: 190,000 - 249,999	0,2026315789

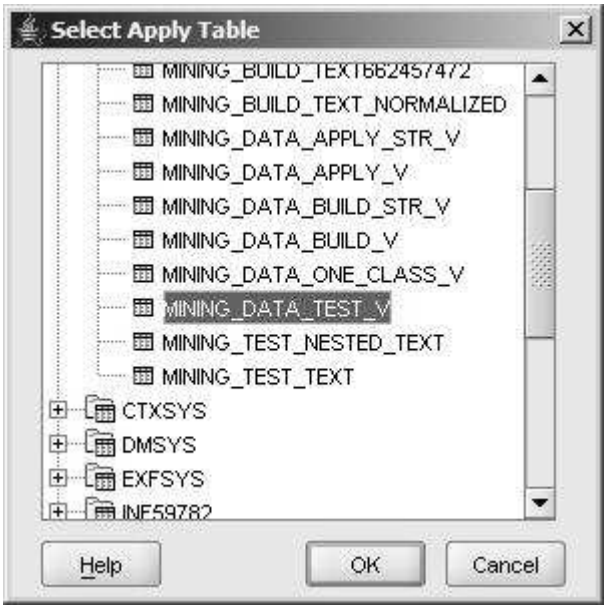
9. Kliknij przycisk Filter. Wskaż wartości graniczne prawdopodobieństwa od 0,5 do 1. Kliknij przycisk OK.



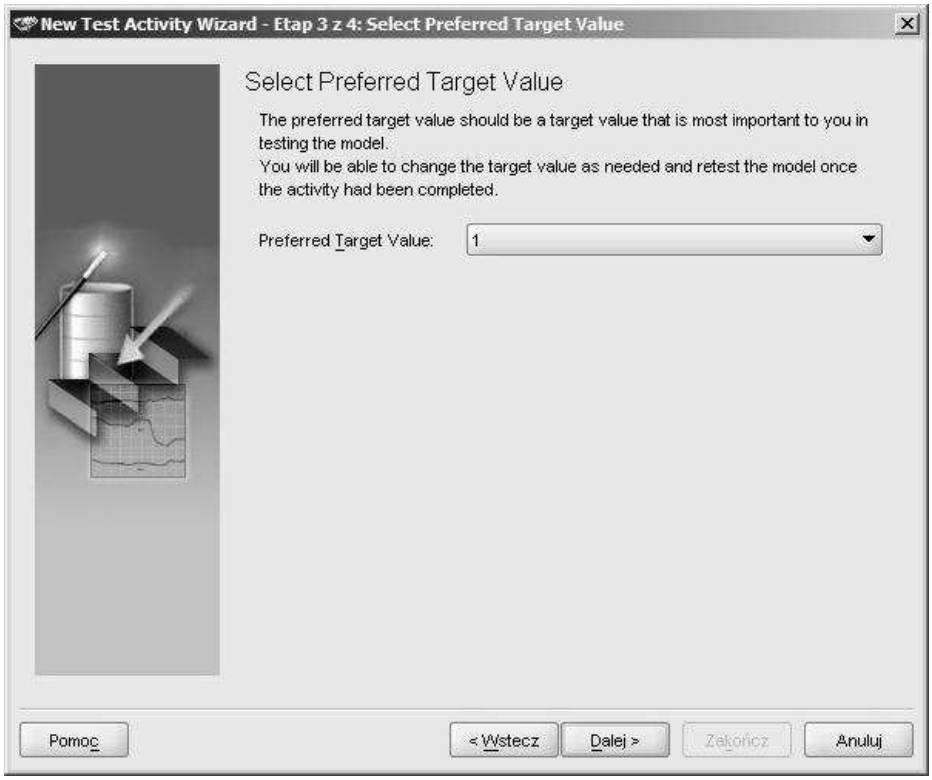
10. Zamknij okno z wynikami budowy klasyfikatora i powróć do głównego okna. Z menu głównego wybierz Activity→Test. Na ekranie powitalnym kliknij przycisk Dalej>.
11. Upewnij się, że zaznaczone jest pole radiowe Build Activity. Rozwiń listę Classification i wybierz model `NB_MINING_DATA_BUILD` jako model do testowania. Kliknij przycisk Dalej>.



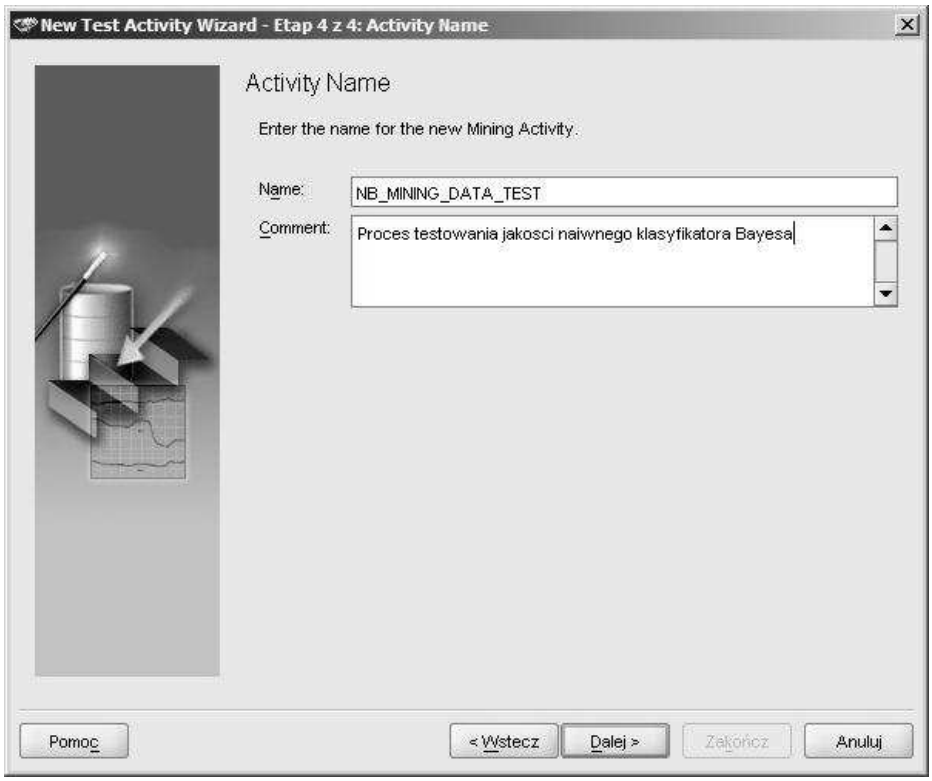
12. Kliknij na odnośnik Select.... Rozwiń węzeł odpowiadający Twojemu schematowi w bazie danych. Jako źródło danych do testowania klasyfikatora wskaż tabelę `MINING_DATA_TEST_V`. Kliknij przycisk OK. Kliknij przycisk Dalej>.



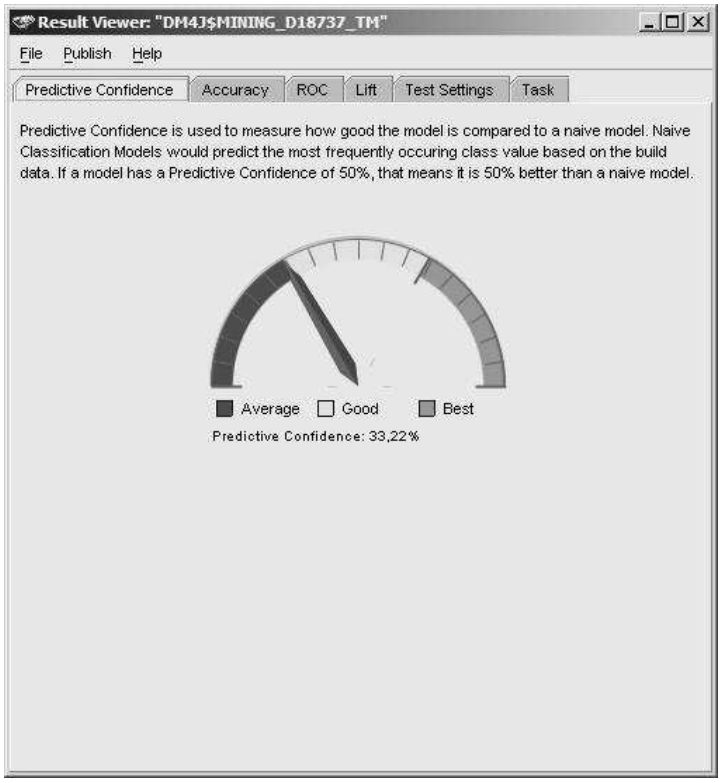
13. Jako preferowaną wartość atrybutu decyzyjnego wybierz 1. Kliknij przycisk Dalej>.



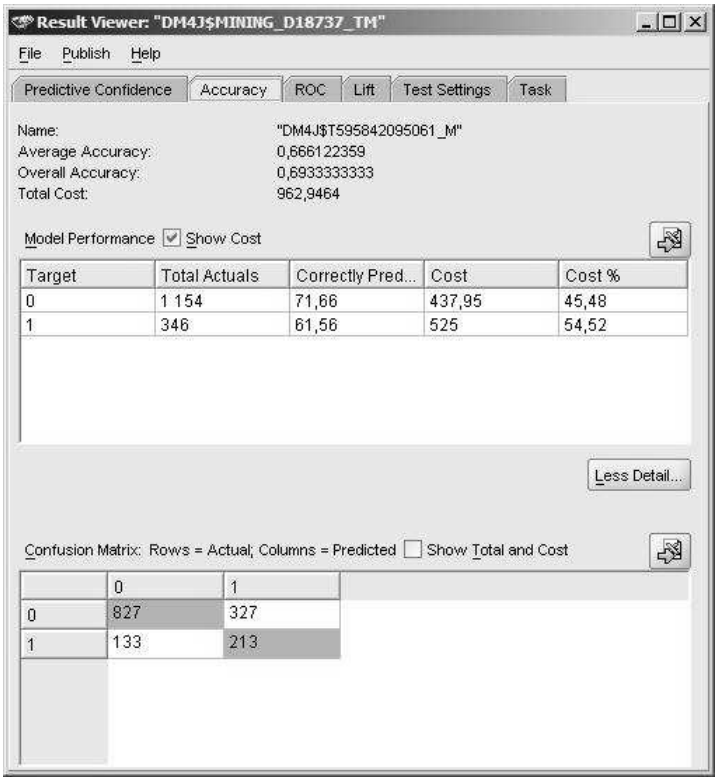
14. Wprowadź nazwę i opis procesu eksploracji. Kliknij przycisk Dalej>. Upewnij się, że zaznaczona jest opcja Run upon finish. Kliknij przycisk Zakończ.



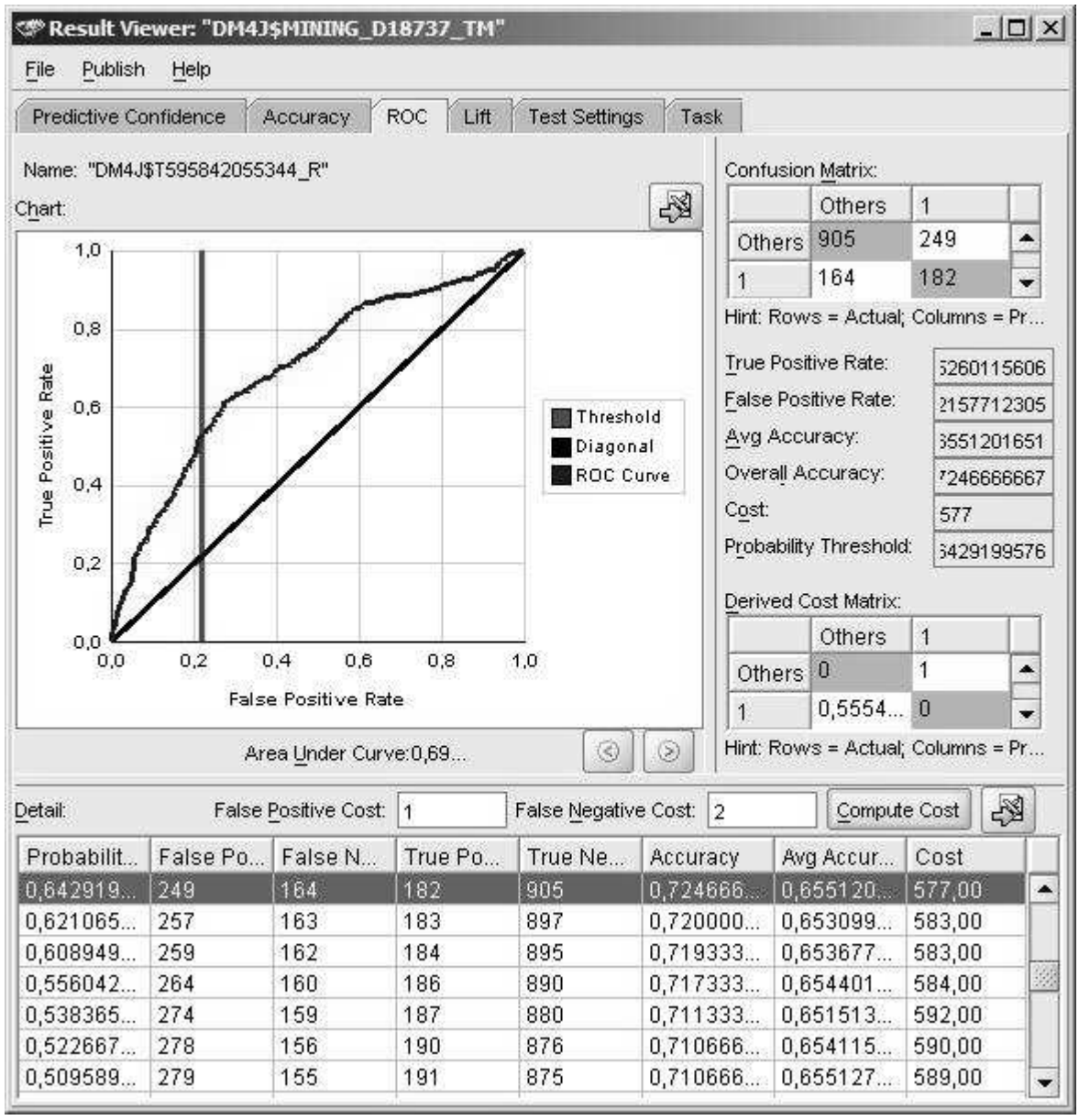
15. Kliknij na odnośnik Result. Na zakładce Predictive Confidence przedstawiona jest dokładność klasyfikatora liczona względem naiwnego klasyfikatora 0-R, który zawsze przewiduje najczęstszą wartość atrybutu decyzyjnego.



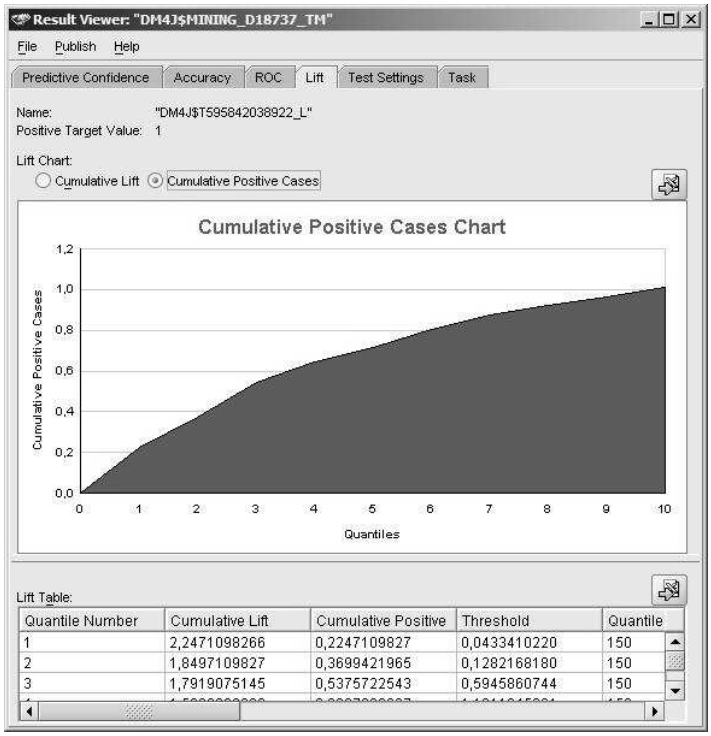
16. Przejdź na zakładkę Accuracy. Zaznacz pole wyboru Show Cost. Kliknij przycisk More Detail.... Przeanalizuj uzyskaną macierz pomyłek.



17. Przejdź na zakładkę ROC. Obejrzyj uzyskaną krzywą Receiver-Operator-Characteristic przedstawiającą stosunek liczby poprawnie sklasyfikowanych instancji (przykładów z wartością atrybutu decyzyjnego 1) do liczby pomyłek (instancji sklasyfikowanych jako należące do klasy 1 podczas gdy w rzeczywistości należą do klasy 0). W dolnej części okna wpisz koszt pomyłki polegającej na niepoprawnym sklasyfikowaniu instancji jako należącej do klasy 1 (False Positive Cost) o wartości 1. Podaj koszt niepoprawnej klasyfikacji instancji jako należącej do klasy 0 (False Negative Cost) o wartości 2 (czyli dwukrotnie większy). Kliknij przycisk Compute Cost. Zobacz, jaką część zbioru testowego należałoby wziąć pod uwagę, aby przy tak zdefiniowanych kosztach pomyłek ogólny koszt błędu klasyfikatora był najmniejszy.

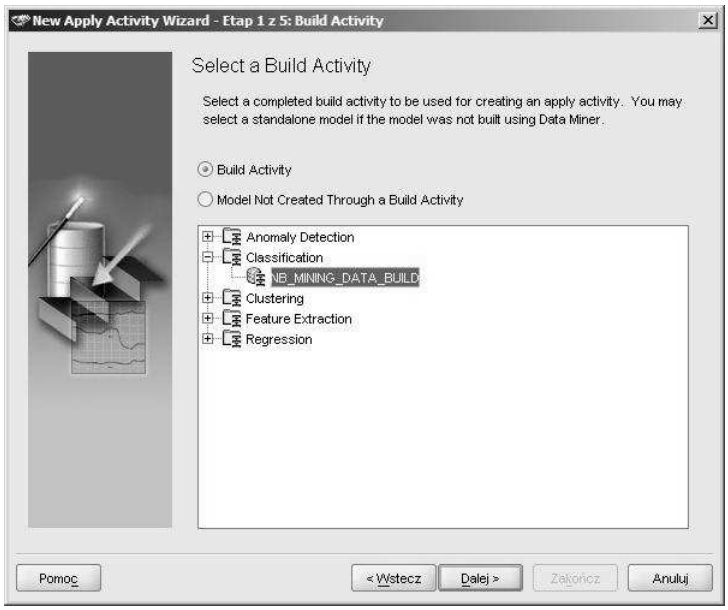


18. Przejdź na zakładkę Lift. Zaznacz pole radiowe Cumulative Positive Cases. Jaki procent zbioru testowego należy rozważyć, aby znaleźć 80% wszystkich instancji należących do klasy 1?

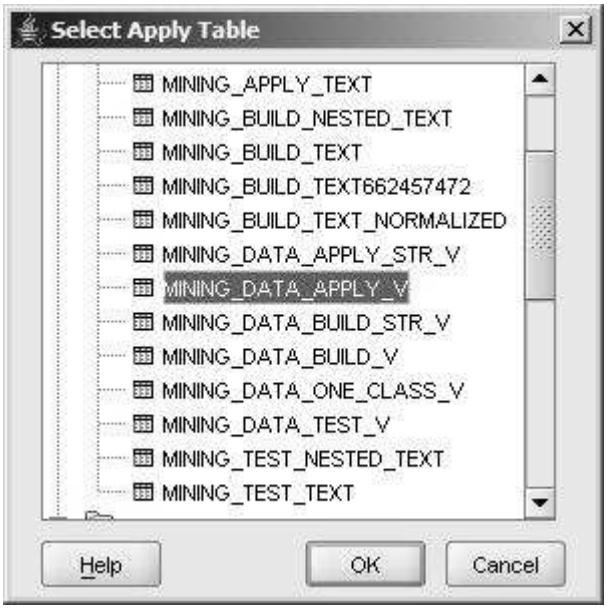


19. Powróć do głównego okna programu. Z menu głównego wybierz Activity→Apply. Na ekranie powitalnym kliknij przycisk Dalej>.

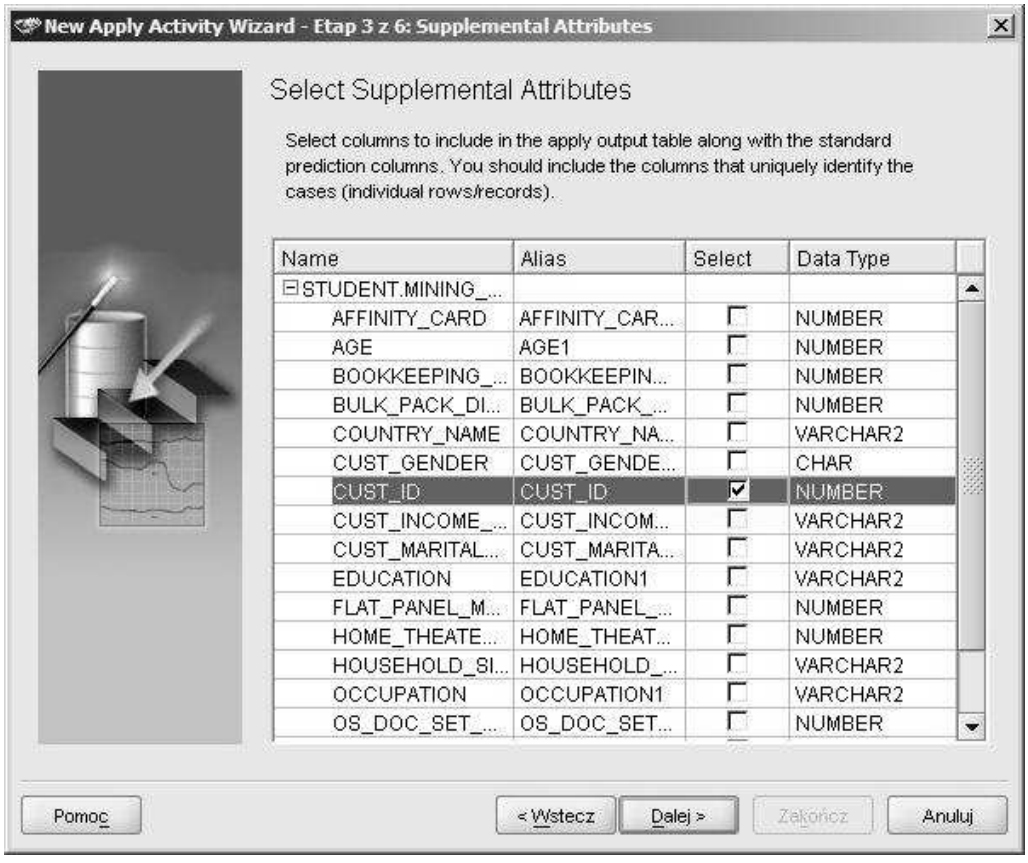
20. Upewnij się, że zaznaczone jest pole radiowe Build Activity. Rozwiń listę Classification i wskaż na model `NB_MINING_DATA_BUILD` jako na model do zastosowania. Kliknij przycisk Dalej>.



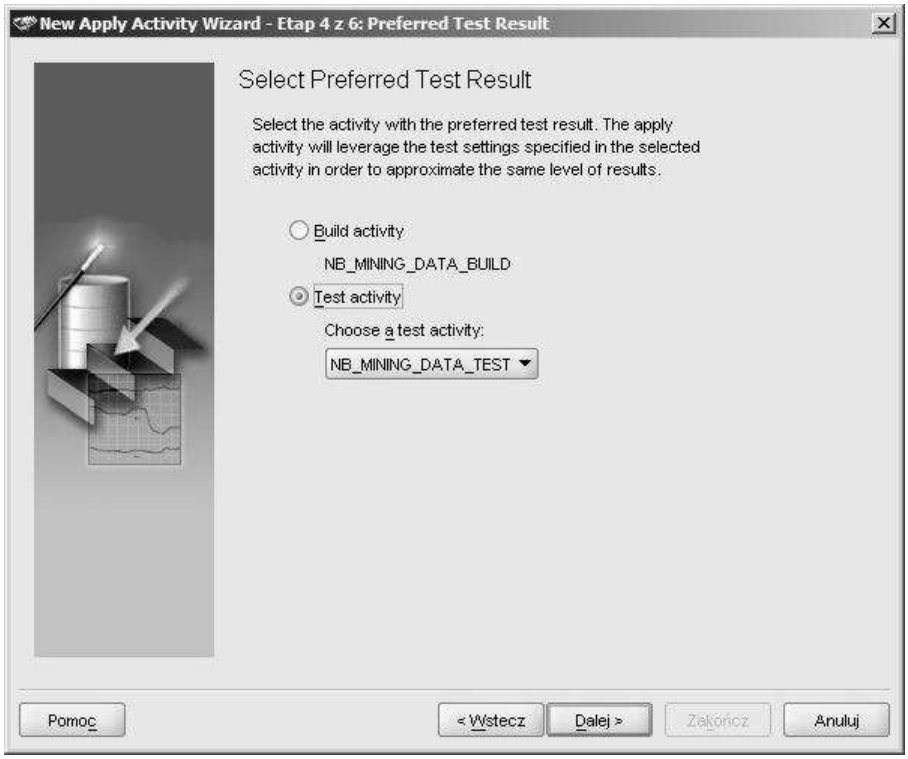
21. Kliknij na odnośnik Select.... Rozwiń węzeł odpowiadający Twojemu schematowi w bazie danych. Jako źródło danych do zastosowania klasyfikatora wskaż tabelę `MINING_DATA_APPLY_V`. Kliknij przycisk OK. Kliknij przycisk Dalej>.



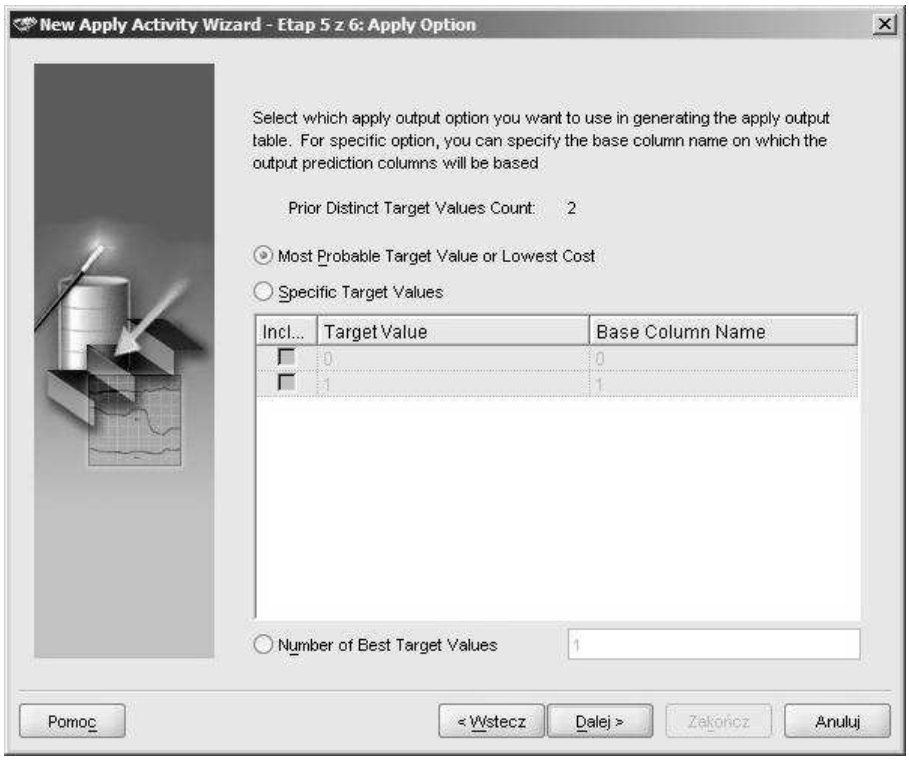
22. Wskaż atrybuty, które powinny się znaleźć w tabeli wynikowej po zastosowaniu klasyfikatora do danych. Upewnij się, że zaznaczony jest klucz podstawowy `CUST_ID`. Kliknij przycisk Dalej>.



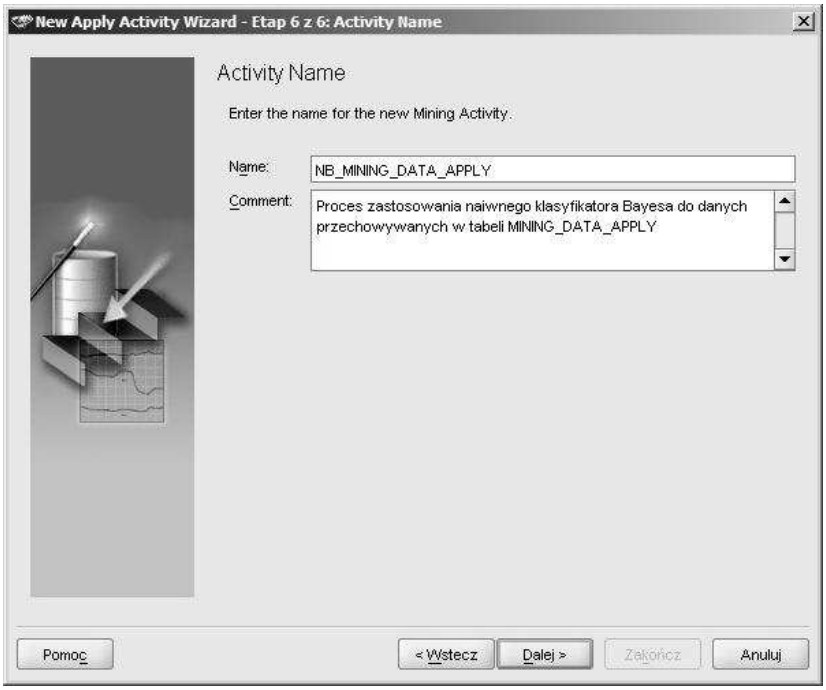
23. Wskaż wykonany wcześniej proces eksploracji zawierający wynik testowania klasyfikatora. Zaznacz pole radiowe Test activity i z listy wybierz proces NB_MINING_DATA_TEST. Kliknij przycisk Dalej>.



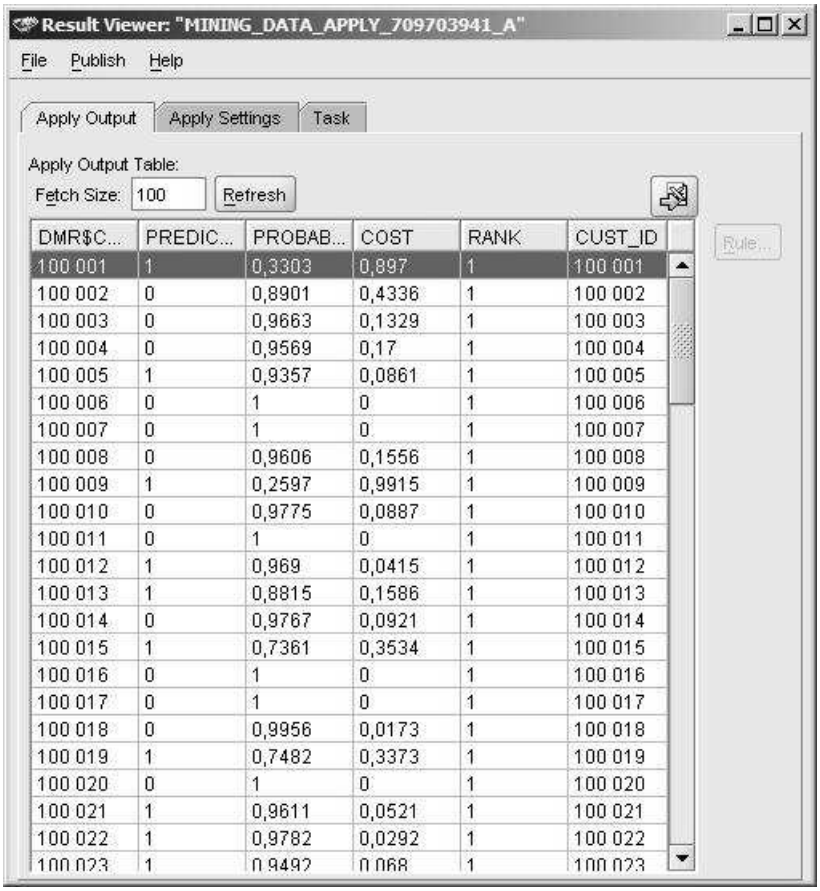
24. Upewnij się, że w kolejnym kroku wybrana jest opcja Most Probable Target Value or Lowest Cost (dla każdej instancji w zbiorze wejściowym zostanie znaleziona jedna najbardziej prawdopodobna wartość atrybutu decyzyjnego). Kliknij przycisk Dalej>.



25. Podaj nazwę i opis procesu eksploracji. Kliknij przycisk Dalej>. Upewnij się, że zaznaczona jest opcja Run upon finish. Kliknij przycisk Zakończ.



26. Kliknij odnośnik Results. Obejrzyj wynik zastosowania klasyfikatora do danych wejściowych. Dla każdej instancji wyświetlone są: przewidywana wartość atrybutu decyzyjnego, prawdopodobieństwo predykcji i koszt związany z predykcją.



DMR$C...	PREDIC...	PROBAB...	COST	RANK	CUST_ID
100 001	1	0,3303	0,897	1	100 001
100 002	0	0,8901	0,4336	1	100 002
100 003	0	0,9663	0,1329	1	100 003
100 004	0	0,9569	0,17	1	100 004
100 005	1	0,9357	0,0861	1	100 005
100 006	0	1	0	1	100 006
100 007	0	1	0	1	100 007
100 008	0	0,9606	0,1556	1	100 008
100 009	1	0,2597	0,9915	1	100 009
100 010	0	0,9775	0,0887	1	100 010
100 011	0	1	0	1	100 011
100 012	1	0,969	0,0415	1	100 012
100 013	1	0,8815	0,1586	1	100 013
100 014	0	0,9767	0,0921	1	100 014
100 015	1	0,7361	0,3534	1	100 015
100 016	0	1	0	1	100 016
100 017	0	1	0	1	100 017
100 018	0	0,9956	0,0173	1	100 018
100 019	1	0,7482	0,3373	1	100 019
100 020	0	1	0	1	100 020
100 021	1	0,9611	0,0521	1	100 021
100 022	1	0,9782	0,0292	1	100 022
100 023	1	0,9492	0,068	1	100 023

Ćwiczenie samodzielne

Na podstawie tabeli `PRACOWNICY` zbuduj perspektywę, która:

- zamieni atrybut `ID_SZEFA` na nazwisko szefa
- doda nowy atrybut `ETAT_SZEFA`
- zamieni atrybut `ZATRUDNIONY` na atrybut numeryczny reprezentujący dekadę zatrudnienia (lata 60-te, 70-te, itd.)
- dokona dyskretyzacji atrybutu `PLACA_POD` na trzy przedziały odpowiadające pensjom niskim, średnim i wysokim
- zamieni atrybut `PLACA_DOD` na binarną flagę 0 (nie otrzymuje dodatków) 1 (otrzymuje dodatki)
- zamieni atrybut `ID_ZESP` na nazwę zespołu

Utworzoną przez siebie perspektywę wykorzystaj do zbudowania naiwnego klasyfikatora Bayesa, które będzie przewidywał wartość atrybutu `ETAT`.
Wykorzystaj poniższy kod do stworzenia tabeli, która posłuży do przetestowania jakości klasyfikatora.

```
CREATE TABLE pracownicy_test AS
  SELECT * FROM pracownicy WHERE 0=1;

INSERT INTO pracownicy_test
(id_prac,nazwisko,etat,id_szefa,zatrudniony,placa_pod,placa_dod,id_zesp)
VALUES (240,'NIEBIESKI','ASYSTENT',130,TO_DATE('01-02-1997','dd-mm-
yyyy'),510,20,20);
INSERT INTO pracownicy_test
(id_prac,nazwisko,etat,id_szefa,zatrudniony,placa_pod,placa_dod,id_zesp)
VALUES (250,'ZOLTY','PROFESOR',100,TO_DATE('01-10-1975','dd-mm-
yyyy'),1110,null,20);
INSERT INTO pracownicy_test
(id_prac,nazwisko,etat,id_szefa,zatrudniony,placa_pod,placa_dod,id_zesp)
VALUES (260,'FIOLETOWY','ADIUNKT',130,TO_DATE('01-03-1984','dd-mm-
yyyy'),580,120,20);
INSERT INTO pracownicy_test
(id_prac,nazwisko,etat,id_szefa,zatrudniony,placa_pod,placa_dod,id_zesp)
VALUES (270,'GRANATOWY','PROFESOR',130,TO_DATE('01-04-1977','dd-mm-
yyyy'),910,60,40);

COMMIT;
```

UWAGA:

- pamiętaj, aby dane testowe poddać identycznym transformacjom jak dane treningowe