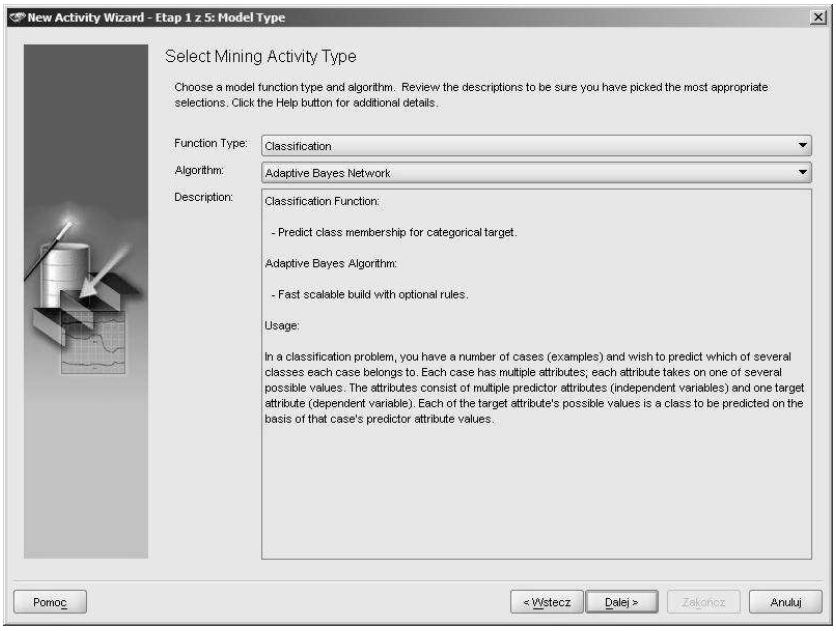
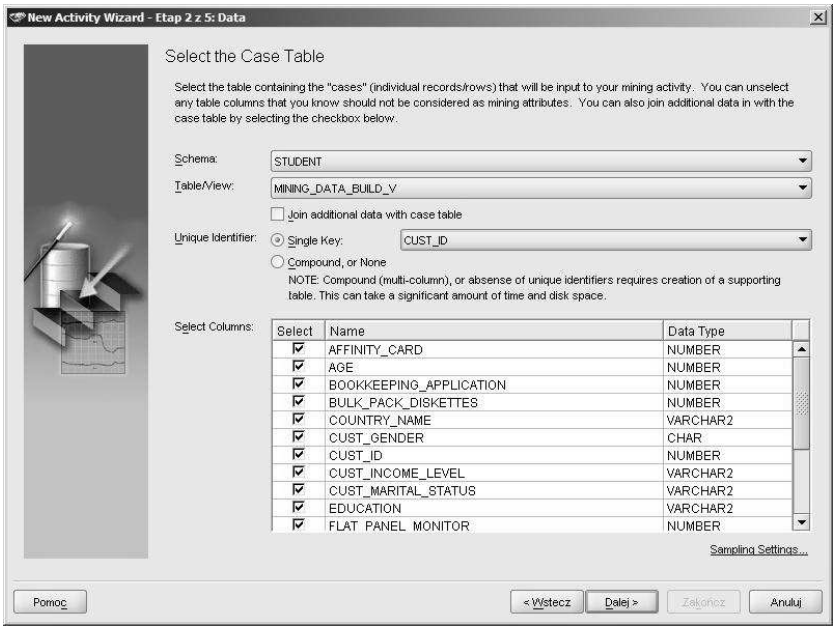
Laboratorium 5

Adaptatywna sieć Bayesa.

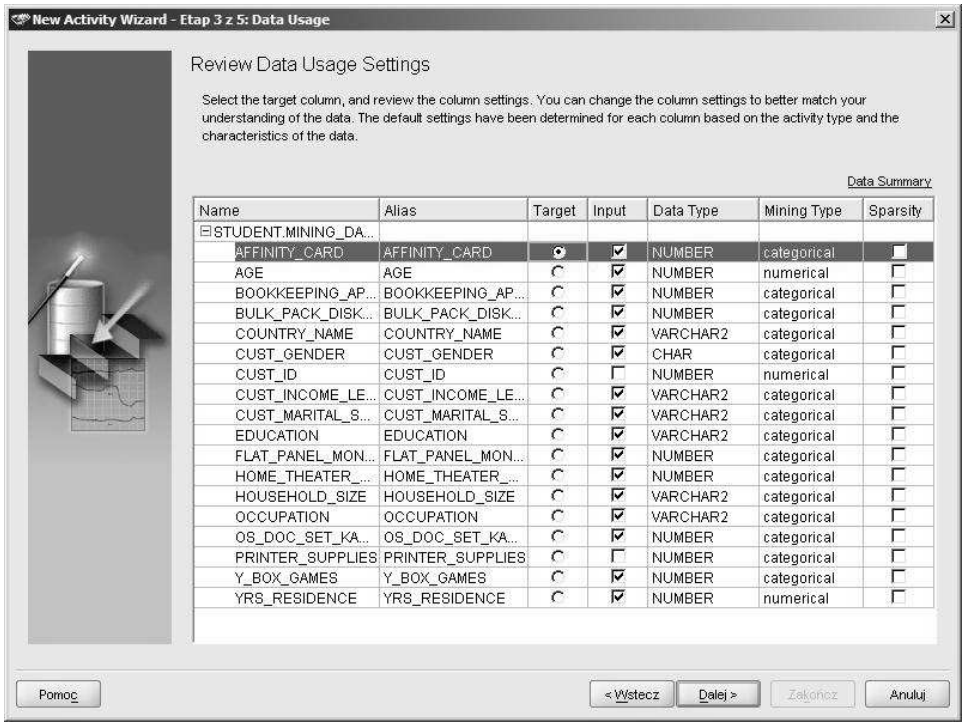
1. Uruchom narzędzie Oracle Data Miner i połącz się z serwerem bazy danych.
2. Z menu głównego wybierz Activity→Build. Na ekranie powitalnym kliknij przycisk Dalej>.
3. Z listy Function Type wybierz Classification. Rozwiń listę Algorithm i wybierz z niej algorytm Adaptive Bayes Network. Kliknij przycisk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



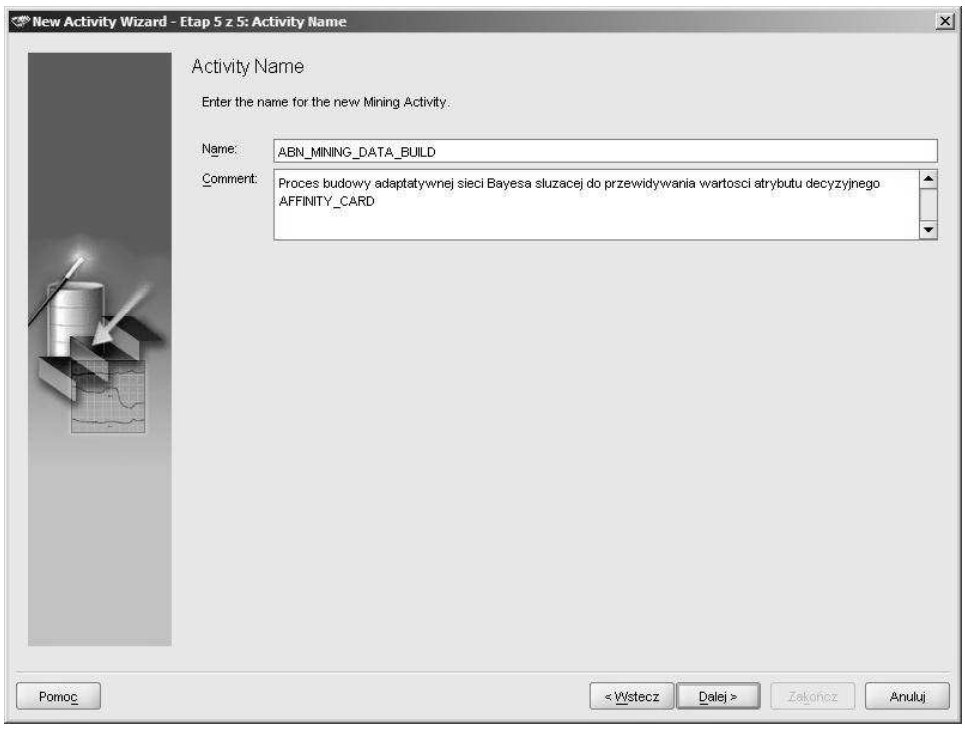
4. Wskaż schemat `STUDENT` i tabelę `MINING_DATA_BUILD_V` jako źródło danych do eksploracji. Jako klucz podstawowy wskaż atrybut `CUST_ID`. Kliknij przycisk Dalej>.



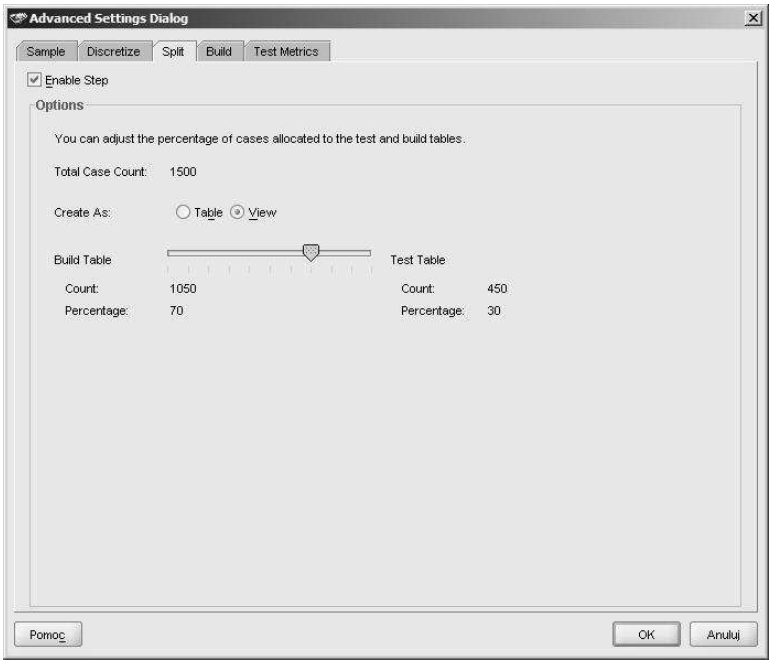
5. Jako atrybut decyzyjny zaznacz atrybut `AFFINITY_CARD` (pole radiowe w kolumnie Target). Upewnij się, że atrybuty `CUST_ID` i `PRINTER_SUPPLIES` są wyłączone z eksploracji (są bezwartościowe i nie niosą żadnej informacji). Kliknij przycisk Dalej>.



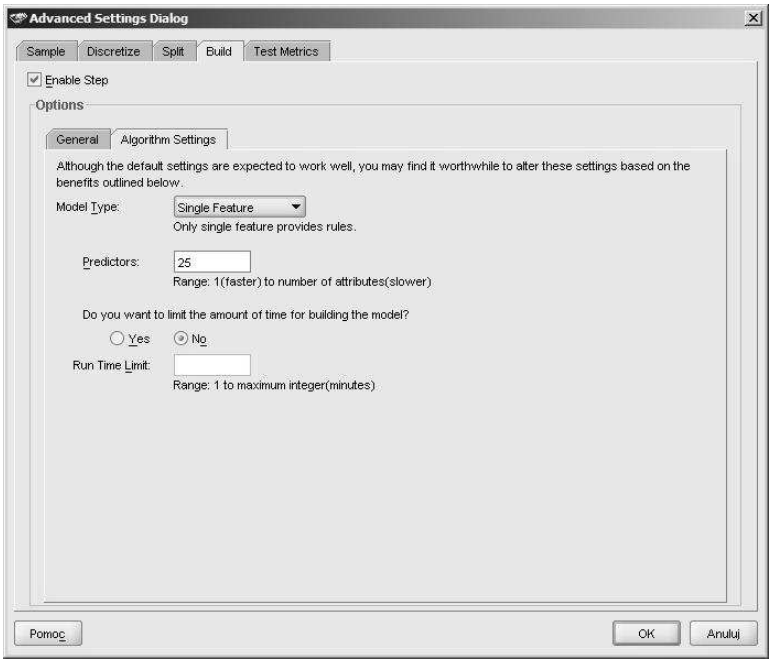
6. Z listy rozwijanej wybierz wartość **1** jako preferowaną wartość atrybutu decyzyjnego (jest to wartość, której poprawne przewidywanie jest najważniejsze, interesuje nas dokładna identyfikacja klientów którzy prawdopodobnie skorzystają z oferowanej im karty lojalnościowej). Kliknij przycisk Dalej>.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



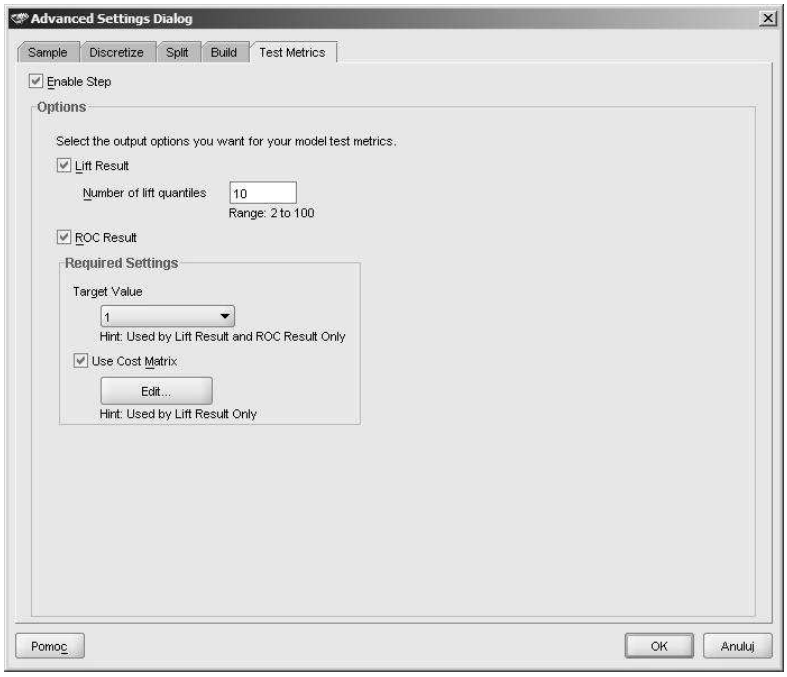
7. Kliknij przycisk Advanced Settings. Upewnij się, że na zakładce Sample opcja próbkowania jest wyłączona (pole wyboru Enable Step jest odznaczone). Przejdź na zakładkę Discretize. Upewnij się, czy automatyczna dyskretyzacja jest włączona (możesz pozostawić domyślne procedury dyskretyzacji). Przejdź na zakładkę Split. Dokonaj podziału zbioru wejściowego na zbiór uczący i testujący w proporcjach 70%-30%, podział powinien wykorzystywać perspektywę.



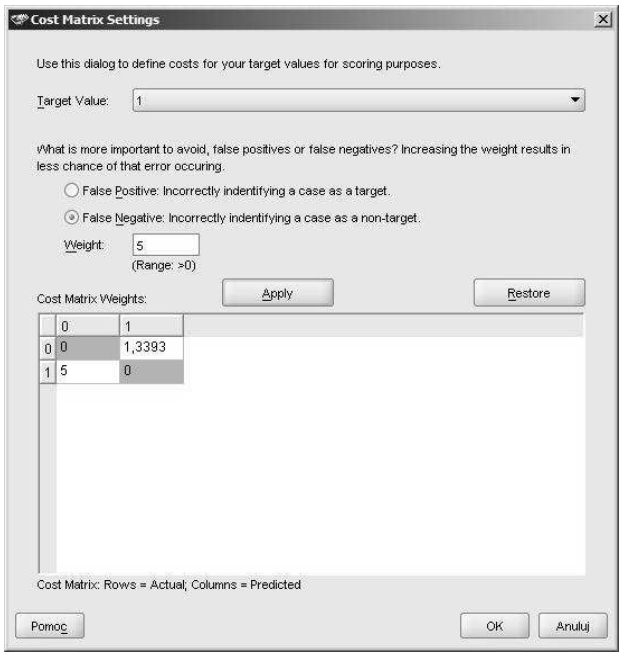
8. Przejdź na zakładkę Build. Upewnij się, że algorytm będzie się starał osiągnąć maksymalną średnią dokładność (w polu Accuracy Goal wybierz opcję Maximum Average Accuracy). Kliknij na zakładkę Algorithm Settings. Jako typ budowanego modelu wskaż Single Feature (to jedyny typ modelu produkujący reguły). Pozostaw domyślną liczbę predykatorów (25) i nie ograniczaj czasowo procesu tworzenia klasyfikatora.



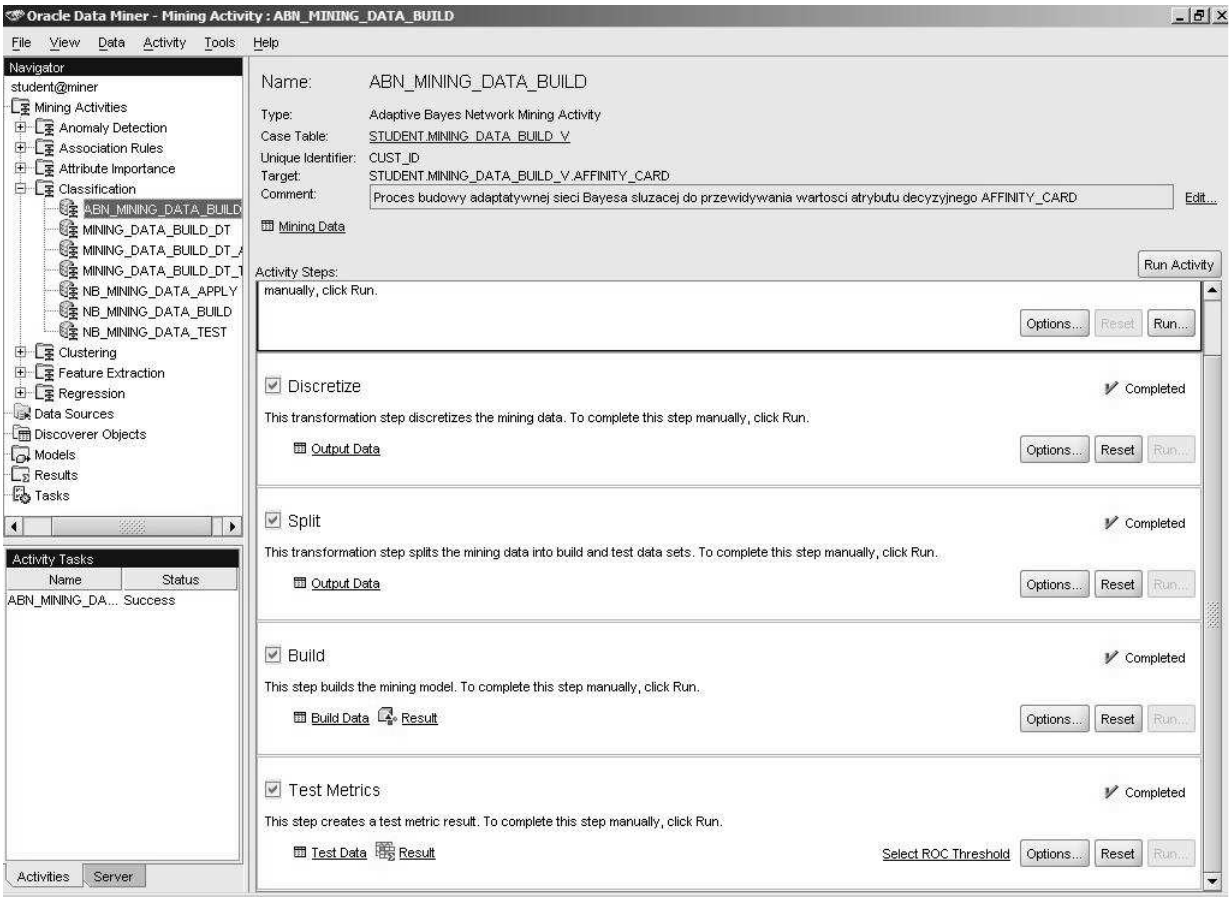
9. Przejdź na zakładkę Test Metrics i upewnij się, że generowanie miar oceny jest włączone (pole Enable Step jest włączone). Pozostaw domyślną liczbę kwantyli dla wykresu krzywej lift. Upewnij się, że włączona jest opcja generowania danych do wykresu Receiver-Operator Characteristic (pole ROC result jest włączone). Jako wartość badaną wskaż wartość 1 (lista rozwijana Target Value). Kliknij przycisk Edit aby zdefiniować macierz kosztów.



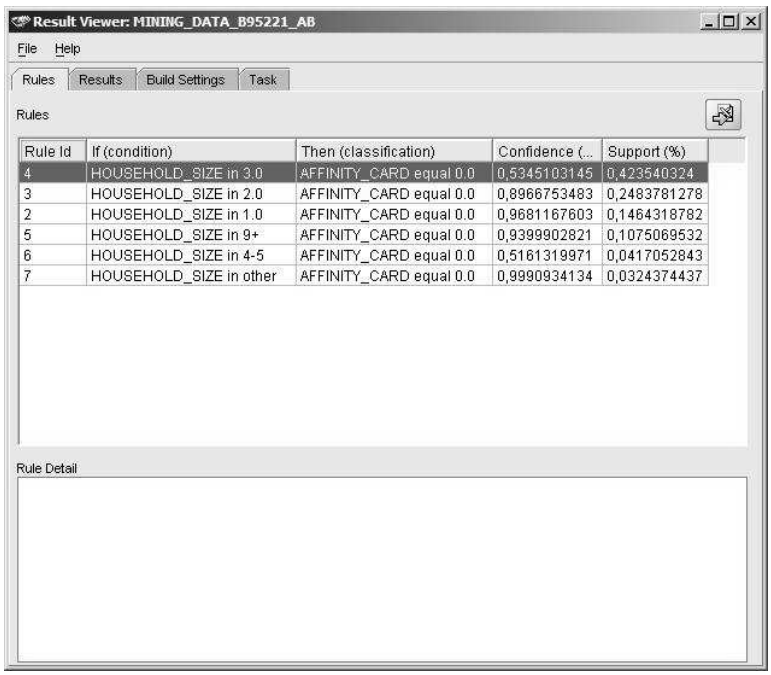
10. Wskaż, że ważniejsze do uniknięcia są błędy klasyfikacji polegające na tym, że osoba potencjalnie zainteresowana kartą lojalnościową (AFFINITY_CARD=1) zostanie niepoprawnie sklasyfikowana jako osoba niezainteresowaną ofertą (taki błąd wiąże się z utratą potencjalnego zysku). Zaznacz pole radiowe False Negative: Incorrectly identifying a case as a non-target. W pole Weight wpisz wartość 5 i kliknij przycisk Apply. Kliknij przycisk OK.



11. Kliknij przycisk OK. Upewnij się, że opcja Run upon finish jest włączona. Kliknij przycisk Zakończ.

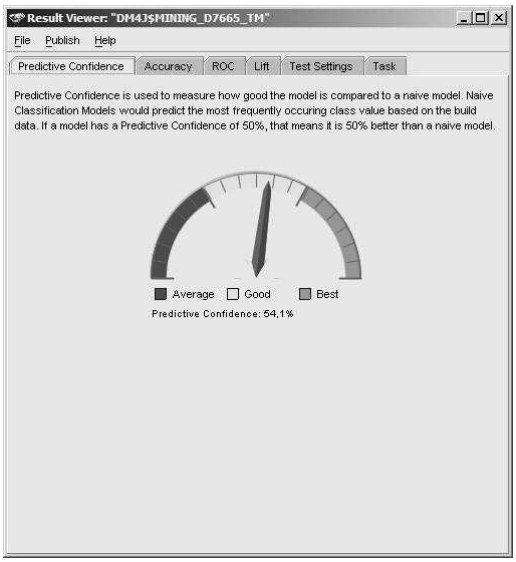


12. Kliknij na odnośnik Result w bloku Build. Jedynym predyktorem okazuje się atrybut HOUSEHOLD_SIZE. Atrybut jest kategoryczny, stąd po jednej regule dla każdej wartości atrybutu.

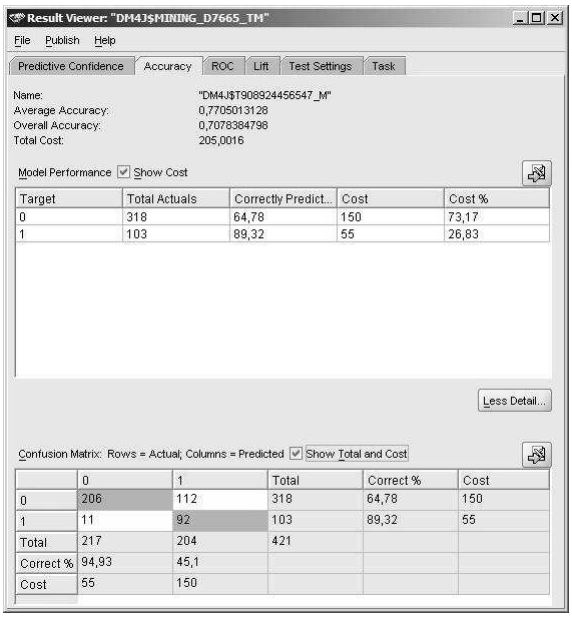


Rule Id	If (condition)	Then (classification)	Confidence (...	Support (%)
4	HOUSEHOLD_SIZE in 3.0	AFFINITY_CARD equal 0.0	0,5345103145	0,423540324
3	HOUSEHOLD_SIZE in 2.0	AFFINITY_CARD equal 0.0	0,8966753483	0,2483781278
2	HOUSEHOLD_SIZE in 1.0	AFFINITY_CARD equal 0.0	0,9681167603	0,1464318782
5	HOUSEHOLD_SIZE in 9+	AFFINITY_CARD equal 0.0	0,9399902821	0,1075069532
6	HOUSEHOLD_SIZE in 4-5	AFFINITY_CARD equal 0.0	0,5161319971	0,0417052843
7	HOUSEHOLD_SIZE in other	AFFINITY_CARD equal 0.0	0,9990934134	0,0324374437

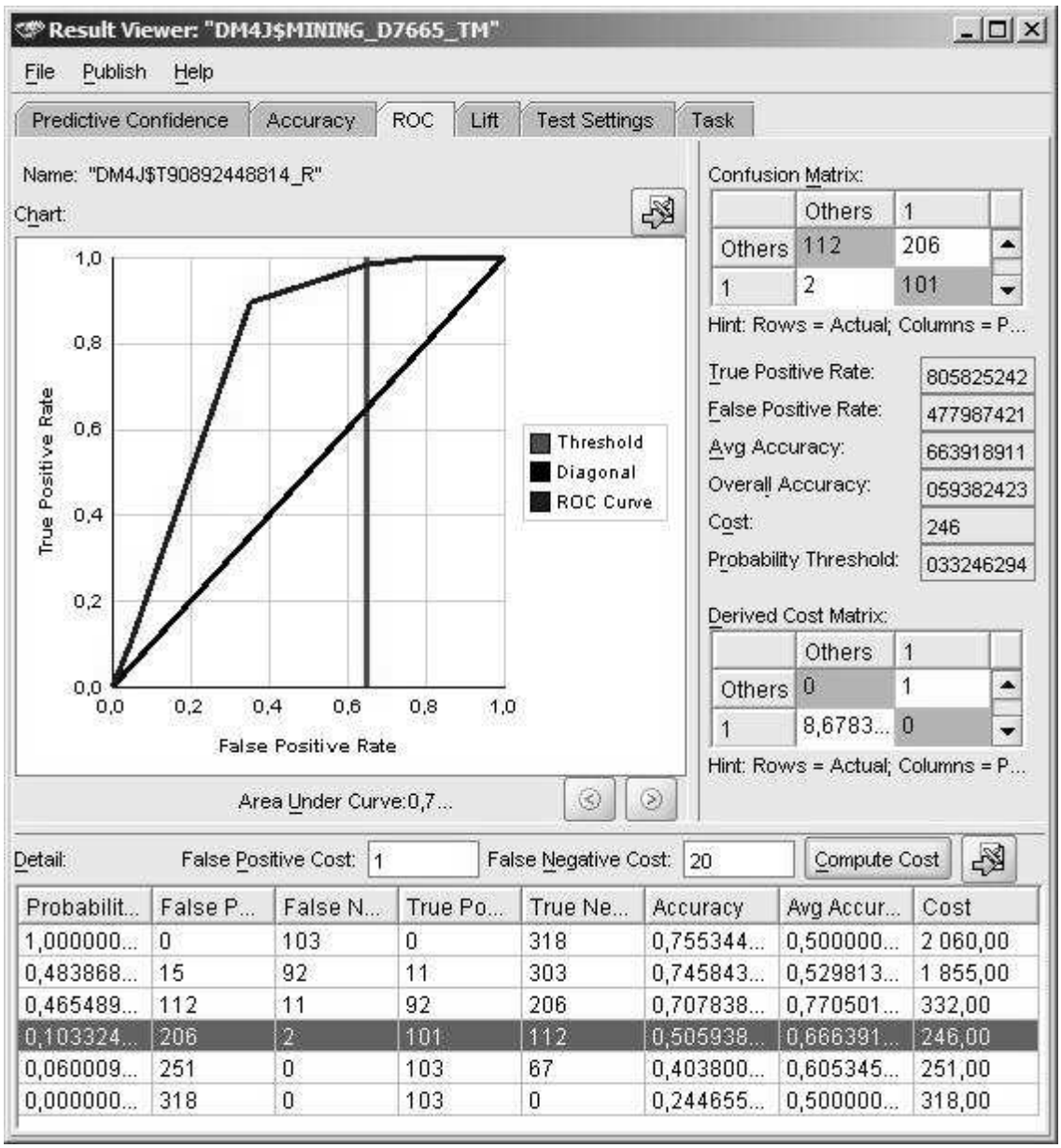
13. Zamknij okno z wynikami budowy klasyfikatora i powróć do głównego okna. Kliknij odnośnik Result w bloku Test Metrics. Na zakładce Predictive Confidence przedstawiona jest dokładność klasyfikatora liczona względem naiwnego klasyfikatora 0-R, który zawsze przewiduje najczęstszą wartość atrybutu decyzyjnego.



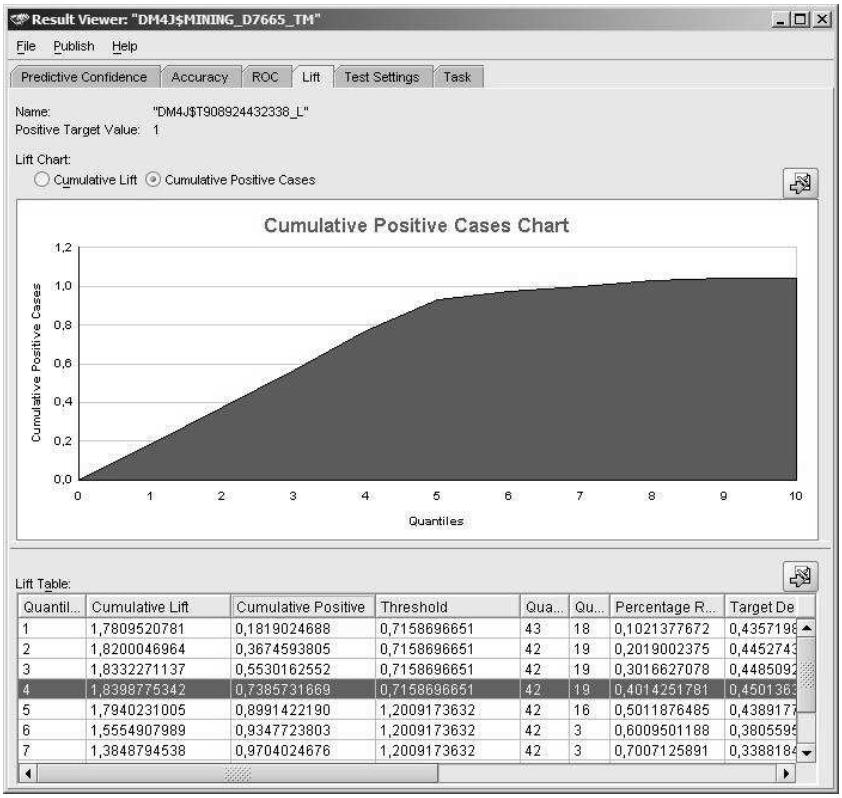
14. Przejdź na zakładkę Accuracy. Zaznacz pole wyboru Show Cost. Kliknij przycisk More Detail... Przeanalizuj uzyskaną macierz pomyłek. Jej interpretacja jest następująca: spośród 318 przypadków należących do klasy 0 w zbiorze testującym prawidłowo przewidziano 206 (64,78%) przypadków, z czym wiązał się koszt pomyłek w wysokości 150 (wyliczony na podstawie macierzy kosztów). Koszt ten stanowi 73,17% ogólnego kosztu błędu klasyfikacji. Spośród 103 przypadków należących do klasy 1 w zbiorze testującym prawidłowo sklasyfikowano 92 (89,32%) przypadki, pozostałe 11 (10,68%) źle sklasyfikowanych przypadków spowodowało koszt w wysokości 55 jednostek. Modyfikacja macierzy kosztów polegająca na zwiększenie kosztu związanego ze sklasyfikowaniem przypadku z klasy 1 jako należącego do klasy na pewno zmniejszyłaby liczbę pomyłek przy klasyfikacji klasy 1, ale gwałtownie zwiększyłaby liczbę pomyłek dotyczących klasyfikacji klasy 0.



15. 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). Znajdź optymalny punkt na krzywej, tzn. punkt o najmniejszym koszcie całkowitym klasyfikacji (kliknij właściwy wiersz w tabeli u dołu okienka). Zmień koszt pomyłki typu False Negative na 20 (koszt pominięcia klienta zainteresowanego kartą lojalnościową jest dwudziestokrotnie większy niż koszt wysłania oferty karty lojalnościowej klientowi niezainteresowanemu kartą). Jaki teraz jest optymalny punkt na krzywej?

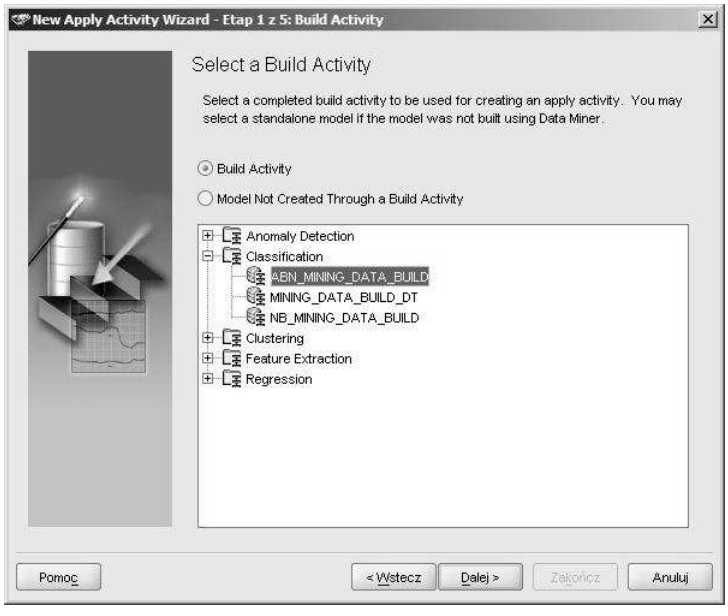


16. Przejdź na zakładkę Lift. Zaznacz pole radiowe Cumulative Positive Cases. Jaki procent zbioru testowego należy rozważyć, aby znaleźć 73% wszystkich instancji należących do klasy 1?

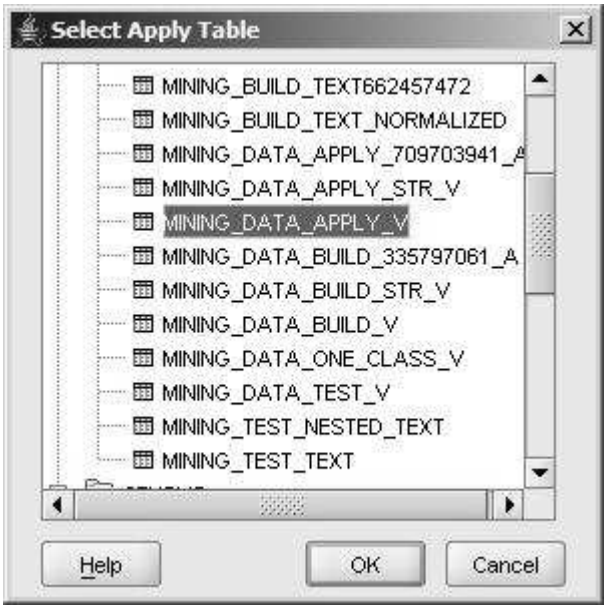


17. Powróć do głównego okna programu. Z menu głównego wybierz Activity→Apply. Na ekranie powitalnym kliknij przycisk Dalej>.

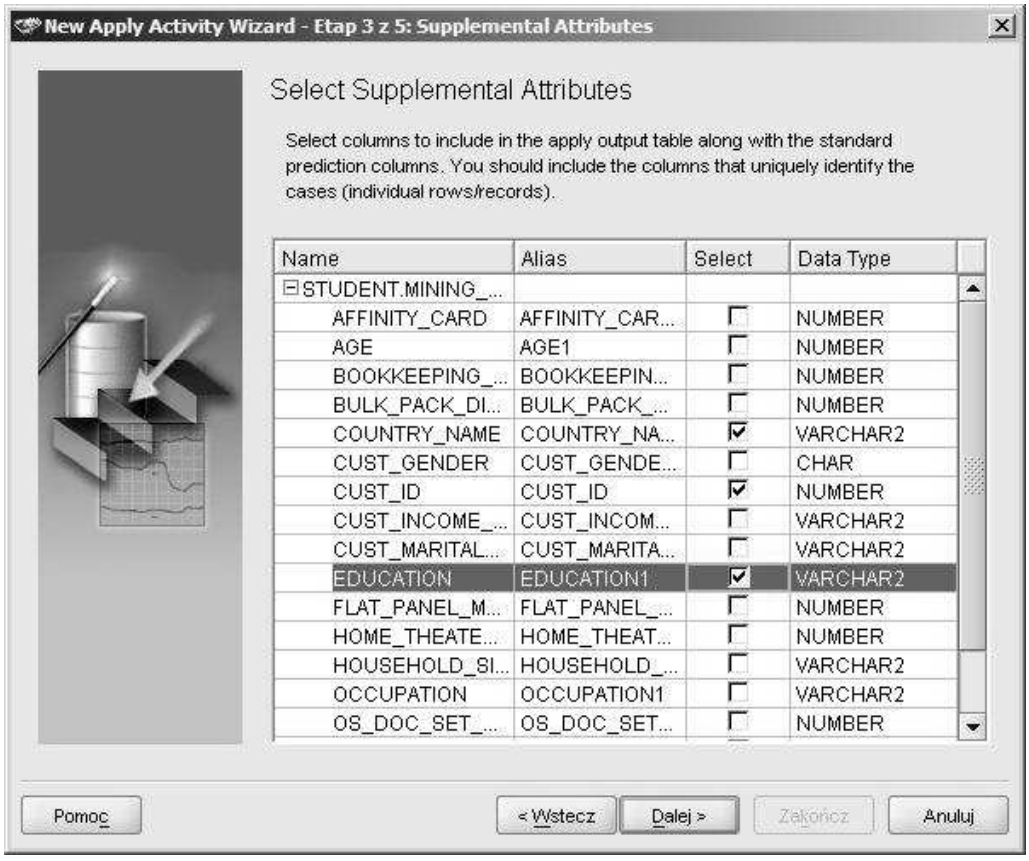
18. Upewnij się, że zaznaczone jest pole radiowe Build Activity. Rozwiń listę Classification i wskaż na model ABN_MINING_DATA_BUILD jako na model do zastosowania. Kliknij przycisk Dalej>.



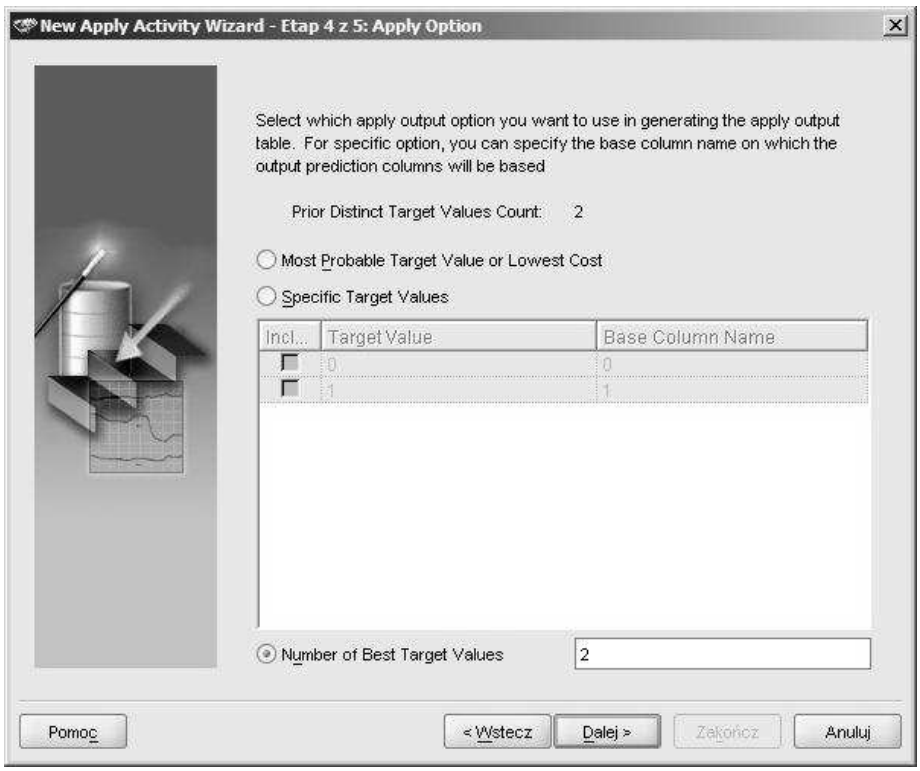
19. Kliknij na odnośnik Select.... Rozwiń węzeł odpowiadający Twojemu schematowi w bazie danych. Jako źródło danych do zastosowania klasyfikatora wskaż tabelę `MINING_DATA_APPLY_V`. Kliknij przycisk OK. Kliknij przycisk Dalej>.



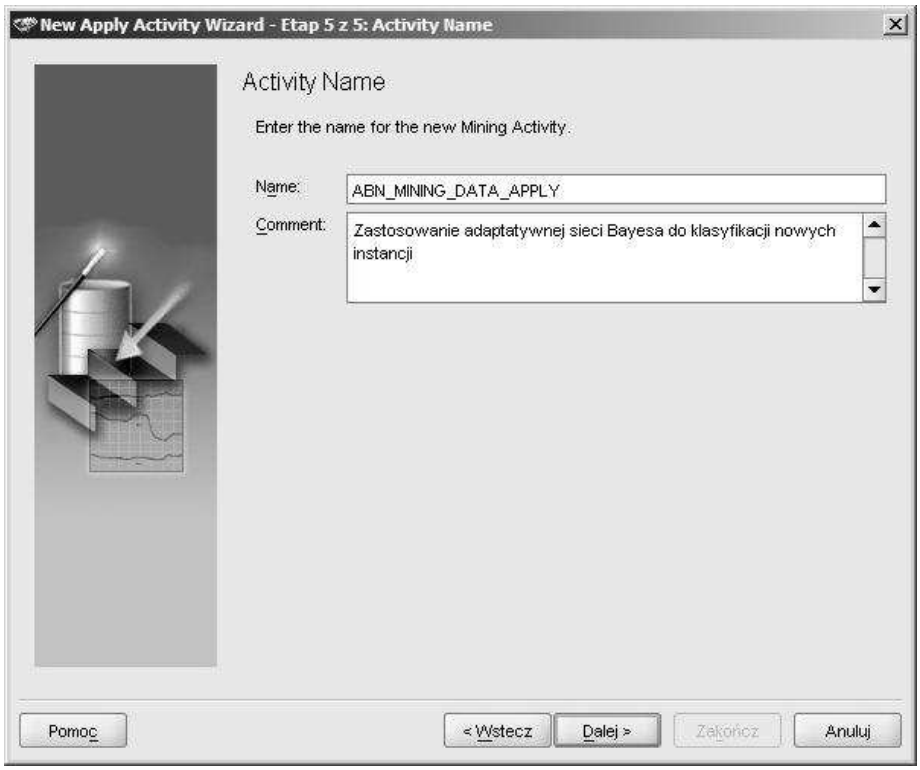
20. Wskaż atrybuty, które powinny się znaleźć w tabeli wynikowej po zastosowaniu klasyfikatora do danych. Upewnij się, że zaznaczony jest klucz podstawowy `CUST_ID` oraz atrybuty `COUNTRY_NAME` i `EDUCATION`. Kliknij przycisk Dalej>.



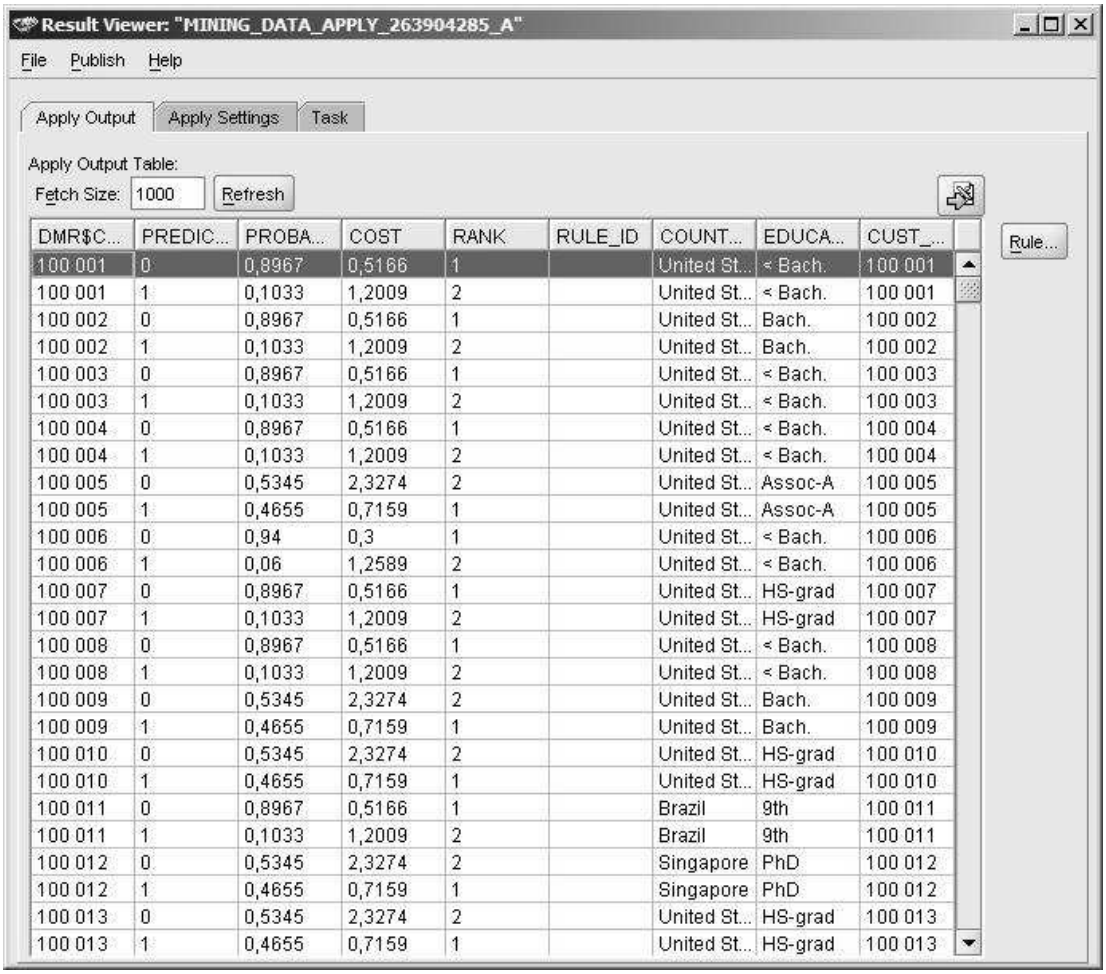
21. Upewnij się, że w kolejnym kroku wybrana jest opcja Number of Best Target Values i wpisz wartośc 2 (dla każdej instancji w zbiorze wejściowym zostaną znalezione dwie najbardziej prawdopodobne wartości atrybutu decyzyjnego). Kliknij przycisk Dalej>.



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



23. Kliknij odnośnik Result. Obejrzyj wynik zastosowania klasyfikatora do danych wejściowych. Zauważ, że dla każdej instancji wyświetlane są dwie możliwości przypisania do klasy decyzyjnej, a każde przypisanie jest opisane prawdopodobieństwem. Przykładowo, klient o identyfikatorze 100 001 należy do klasy 0 z prawdopodobieństwem 89,67% i do klasy 1 z prawdopodobieństwem 10,33%.



DMR$C...	PREDIC...	PROBA...	COST	RANK	RULE_ID	COUNT...	EDUCA...	CUST_...
100 001	0	0,8967	0,5166	1		United St...	< Bach.	100 001
100 001	1	0,1033	1,2009	2		United St...	< Bach.	100 001
100 002	0	0,8967	0,5166	1		United St...	Bach.	100 002
100 002	1	0,1033	1,2009	2		United St...	Bach.	100 002
100 003	0	0,8967	0,5166	1		United St...	< Bach.	100 003
100 003	1	0,1033	1,2009	2		United St...	< Bach.	100 003
100 004	0	0,8967	0,5166	1		United St...	< Bach.	100 004
100 004	1	0,1033	1,2009	2		United St...	< Bach.	100 004
100 005	0	0,5345	2,3274	2		United St...	Assoc-A	100 005
100 005	1	0,4655	0,7159	1		United St...	Assoc-A	100 005
100 006	0	0,94	0,3	1		United St...	< Bach.	100 006
100 006	1	0,06	1,2589	2		United St...	< Bach.	100 006
100 007	0	0,8967	0,5166	1		United St...	HS-grad	100 007
100 007	1	0,1033	1,2009	2		United St...	HS-grad	100 007
100 008	0	0,8967	0,5166	1		United St...	< Bach.	100 008
100 008	1	0,1033	1,2009	2		United St...	< Bach.	100 008
100 009	0	0,5345	2,3274	2		United St...	Bach.	100 009
100 009	1	0,4655	0,7159	1		United St...	Bach.	100 009
100 010	0	0,5345	2,3274	2		United St...	HS-grad	100 010
100 010	1	0,4655	0,7159	1		United St...	HS-grad	100 010
100 011	0	0,8967	0,5166	1		Brazil	9th	100 011
100 011	1	0,1033	1,2009	2		Brazil	9th	100 011
100 012	0	0,5345	2,3274	2		Singapore	PhD	100 012
100 012	1	0,4655	0,7159	1		Singapore	PhD	100 012
100 013	0	0,5345	2,3274	2		United St...	HS-grad	100 013
100 013	1	0,4655	0,7159	1		United St...	HS-grad	100 013

Ćwiczenie samodzielne

Powtórz ćwiczenie dotyczące naiwnego klasyfikatora Bayesa i porównaj jakość uzyskanych wyników. 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 adaptatywnej sieci Bayesa, która będzie przewidywać 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