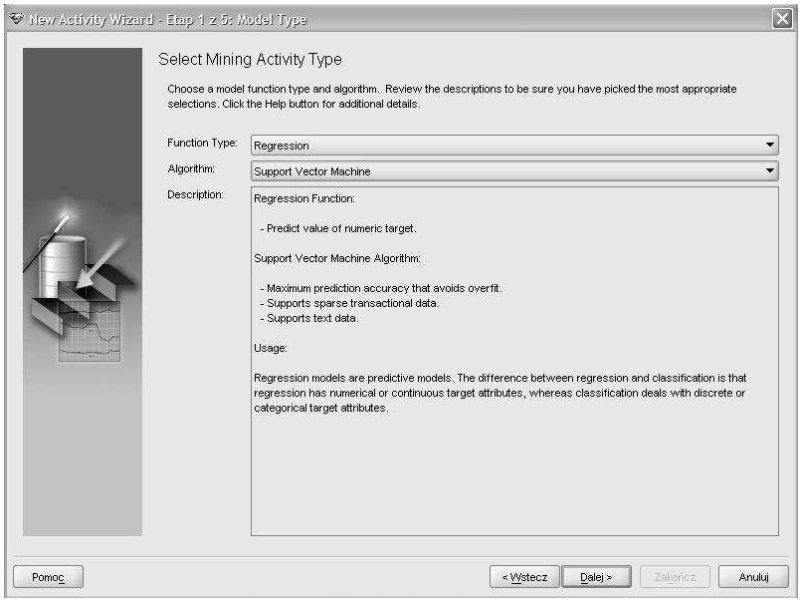
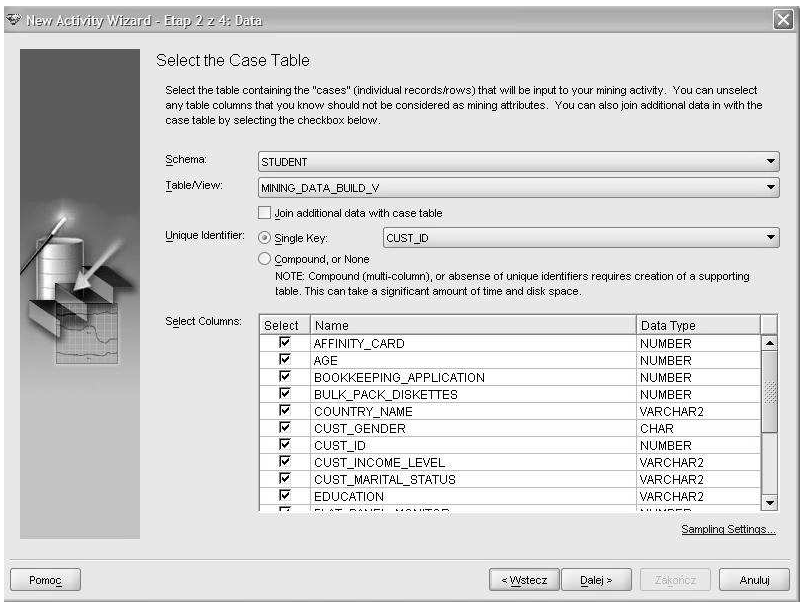
# Laboratorium 11

## *Regresja SVM.*

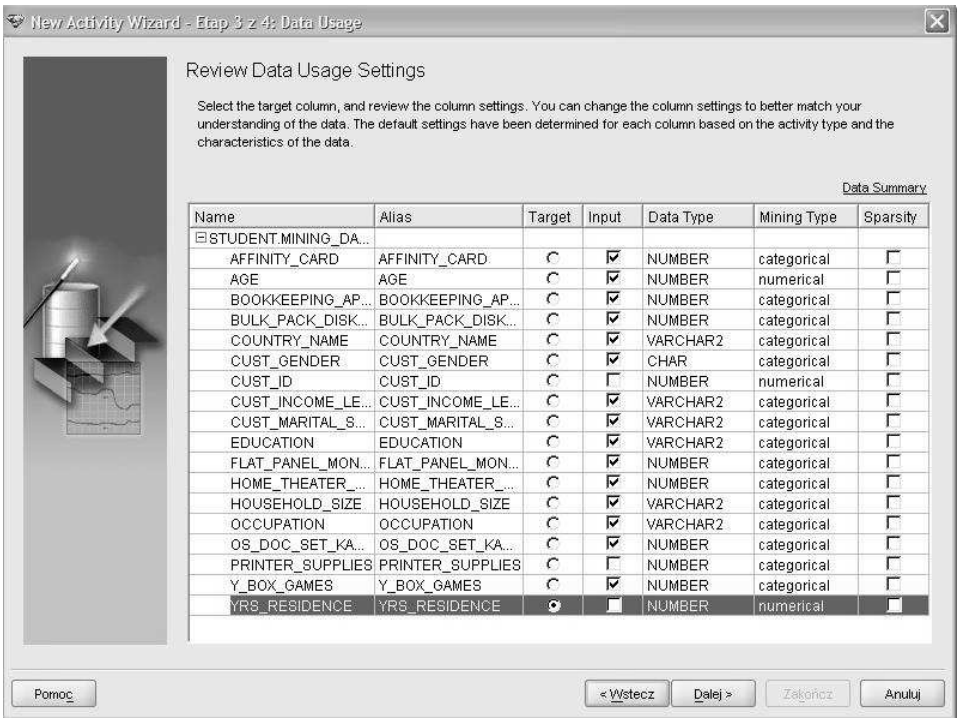
1. Uruchom narzędzie Oracle Data Miner i połącz się z serwerem bazy danych.
2. Z menu głównego wybierz Activity→Build. Na ekranie powitalnym kliknij przycisk Dalej>.
3. Z listy Function Type wybierz Regression. Rozwiń listę Algorithm i wybierz z niej algorytm Support Vector Machines. Kliknij przycisk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



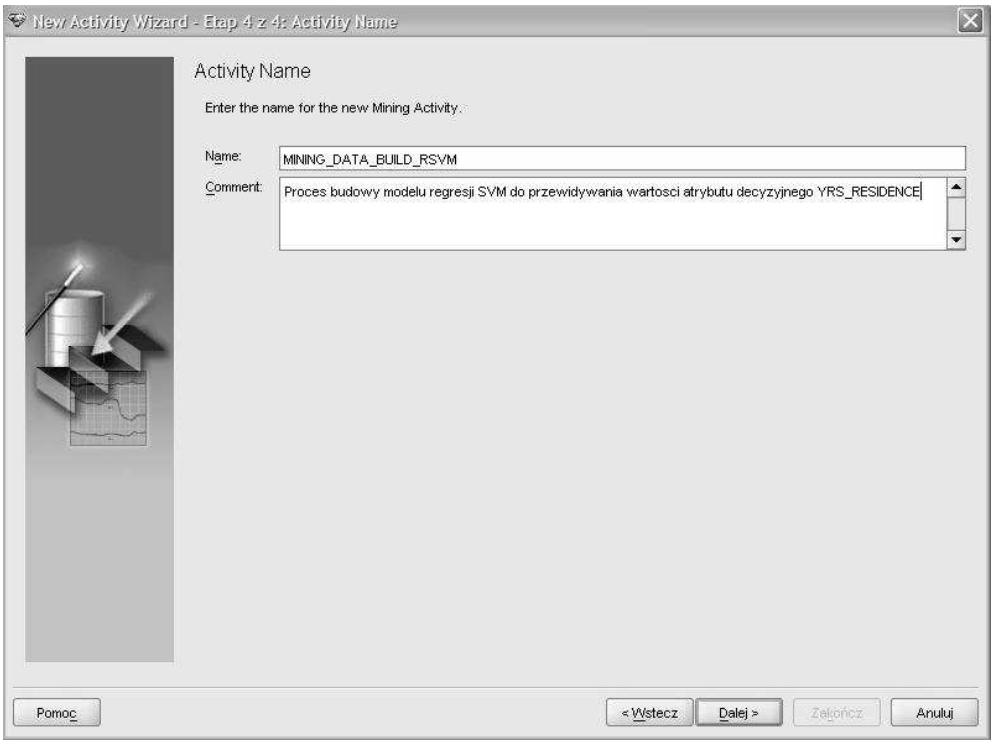

4. Wskaż schemat STUDENT i tabelę MINING_DATA_BUILD_V jako źródło danych do eksploracji. Jako klucz podstawowy wskaż atrybut CUST_ID. Kliknij przycisk Dalej>.



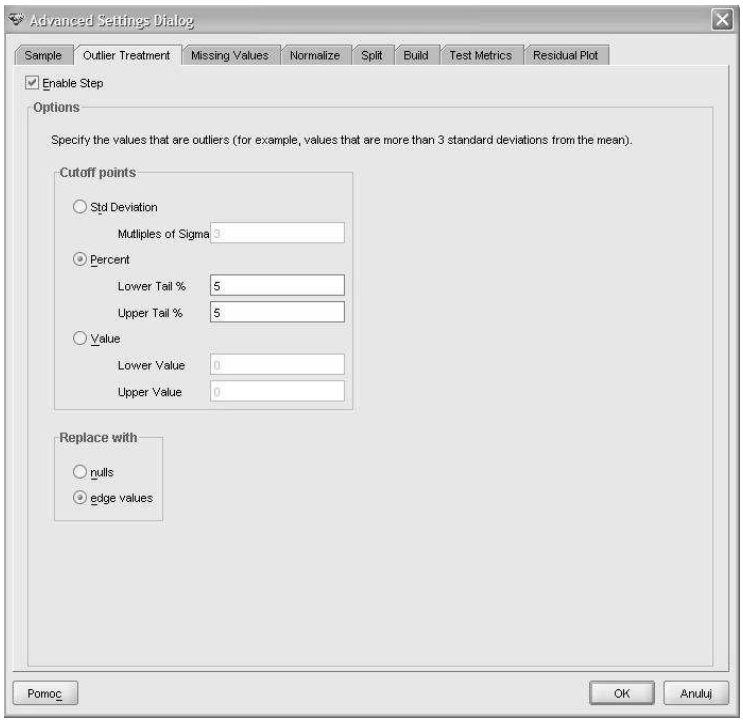
5. Jako atrybut decyzyjny zaznacz atrybut YRS_RESIDENCE (pole radiowe w kolumnie Target). Zwróć uwagę, aby wartość atrybutu decyzyjnego została wyłączona z budowy klasyfikatora (pole wyboru Input dla atrybutu YRS_RESIDENCE musi być odznaczone). Upewnij się, że atrybuty CUST_ID i PRINTER_SUPPLIES są wyłączone z eksploracji (są bezwartościowe i nie niosą żadnej informacji). Kliknij przycisk Dalej>.



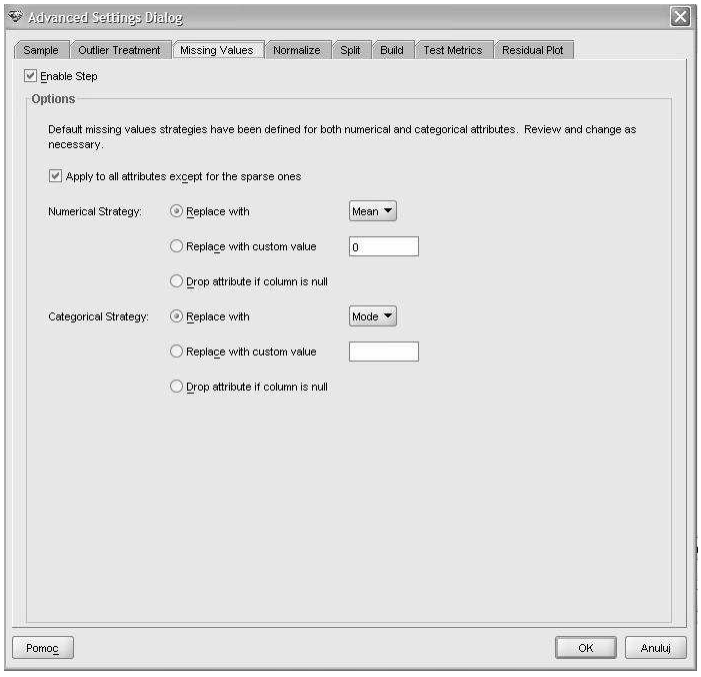

6. Podaj nazwę i krótki opis procesu eksploracji. Kliknij przycisk Dalej>.



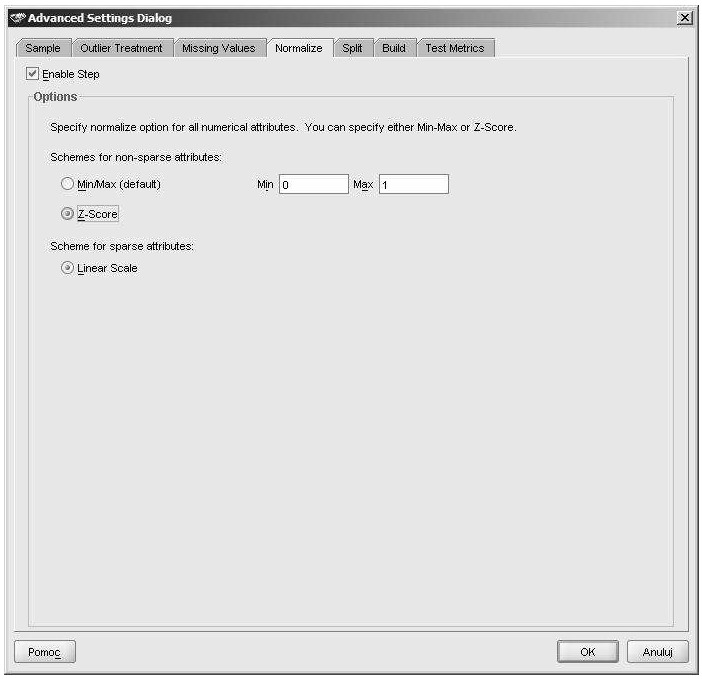
7. Kliknij przycisk Advanced Settings. Upewnij się, że na zakładce Sample opcja próbkowania jest wyłączona (pole wyboru Enable Step jest odznaczone). Przejdź na zakładkę Outlier Treatment. Algorytm SVM jest bardzo czuły na występowanie osobliwości. Oznacz jako osobliwości po 5% wartości z każdego końca przedziału wartości zastępując usuwane osobliwości wartościami brzegowymi.



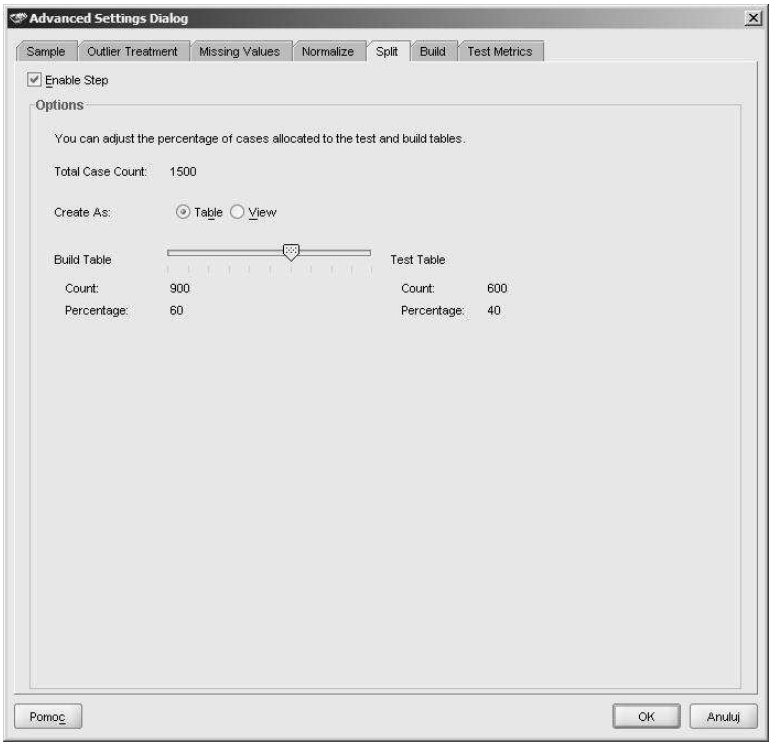

8. Przejdź na zakładkę Missing Values. Upewnij się, że przetwarzanie brakujących wartości jest włączone (pole wyboru Enable Step musi być zaznaczone). Wartości puste występujące w atrybutach numerycznych zamień na wartość średnią (Mean), a wartości puste występujące w atrybutach kategorycznych zamień na wartość modalną (Mode).



9. Przejdź na zakładkę Normalize. Algorytm SVM wymaga, aby wszystkie atrybuty numeryczne były znormalizowane. Jako metodę normalizacji wybierz wyrażenie wartości w liczbie odchyleń standardowych od średniej (zaznacz pole radiowe Z-Score).

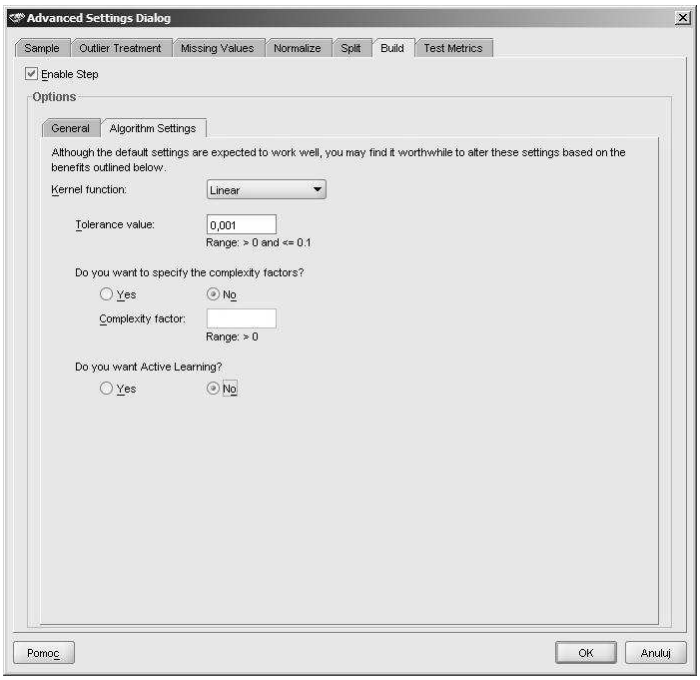



10. Przejdź na zakładkę Split. Dokonaj podziału zbioru wejściowego na zbiór uczący i testujący w proporcjach 60%-40%, podział powinien wykorzystywać tabelę.

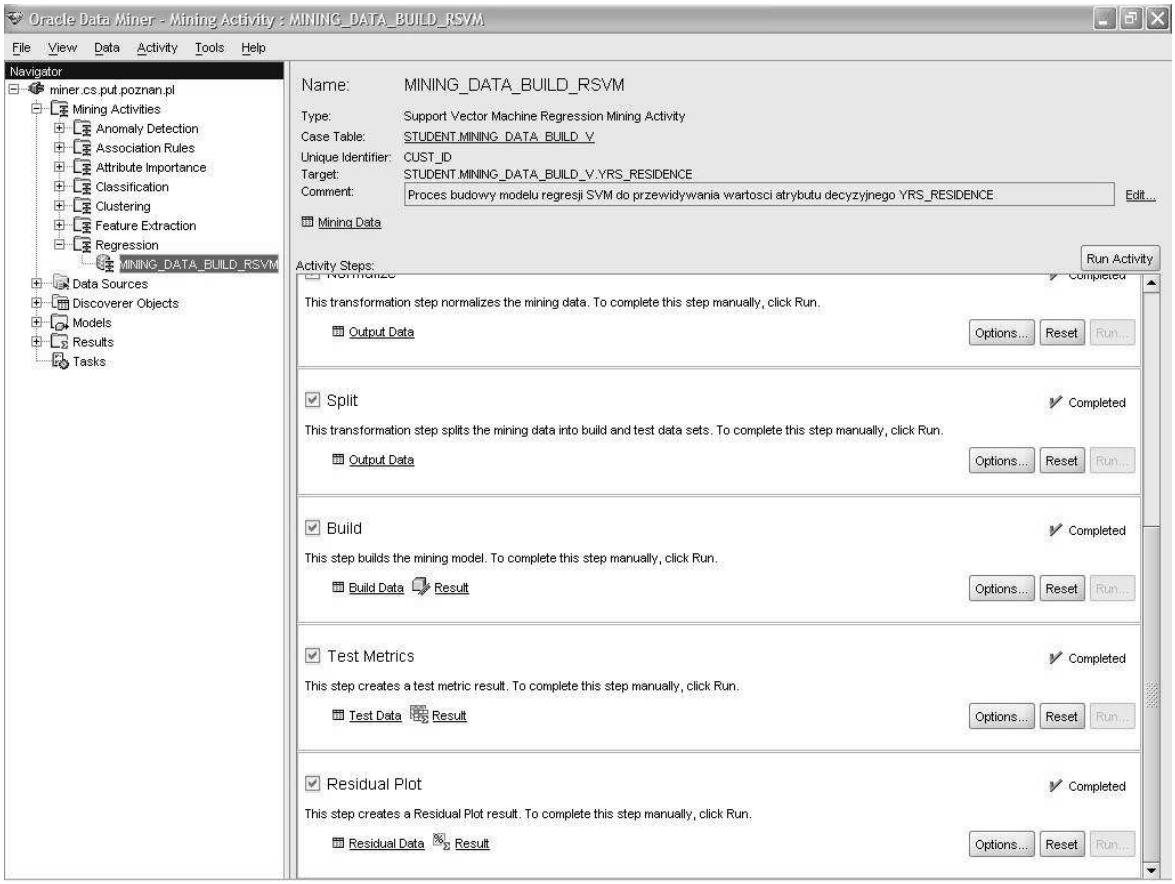



11. Przejdź na zakładkę Build. Upewnij się, że algorytm będzie się starał osiągnąć maksymalną średnią dokładność (w polu Accuracy Goal wybierz opcję Maximum Average Accuracy). Kliknij na zakładkę Algorithm Settings. Jako rodzaj funkcji

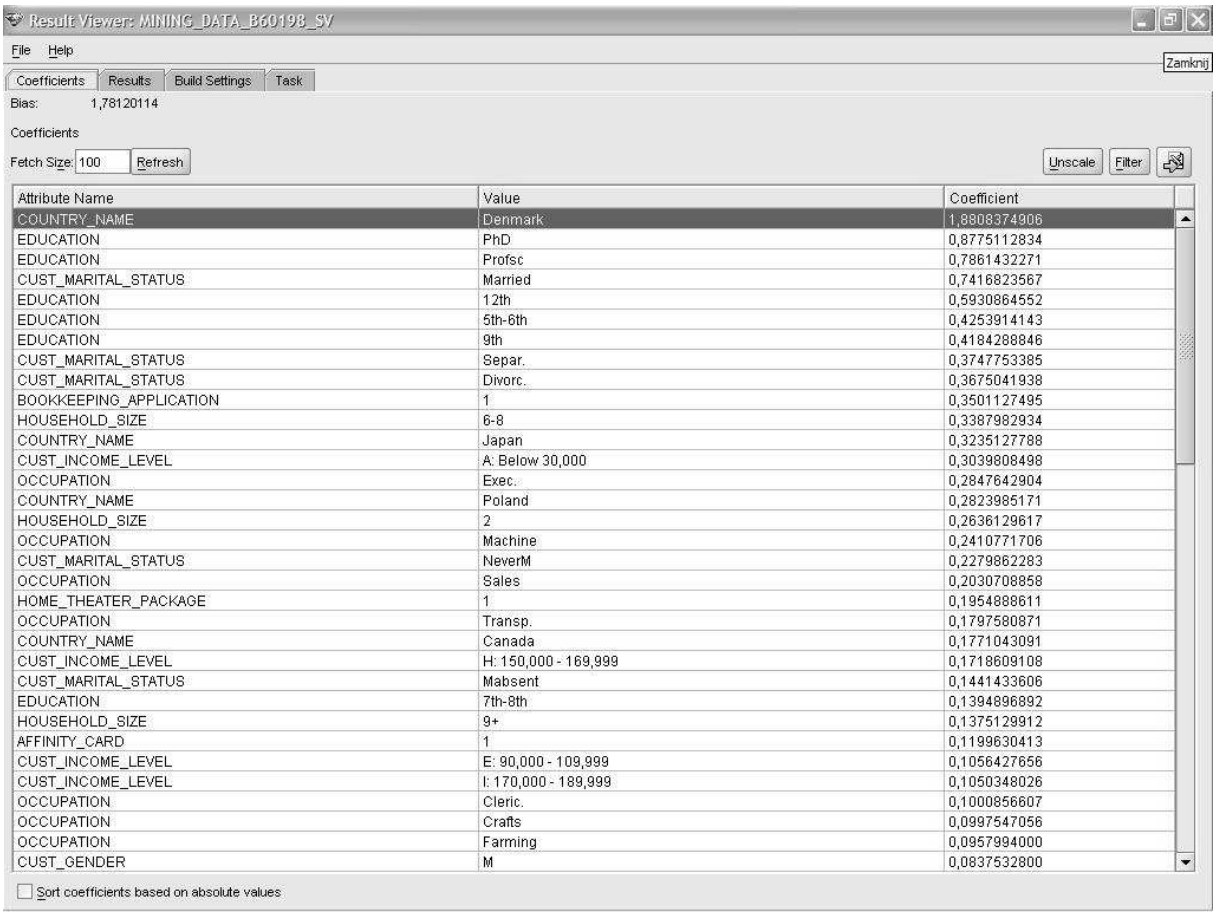
jądrowej wskaż funkcję liniową. Koniecznie wyłącz opcję aktywnego uczenia (pole radiowe Do you want Active Learning?, opcja No).



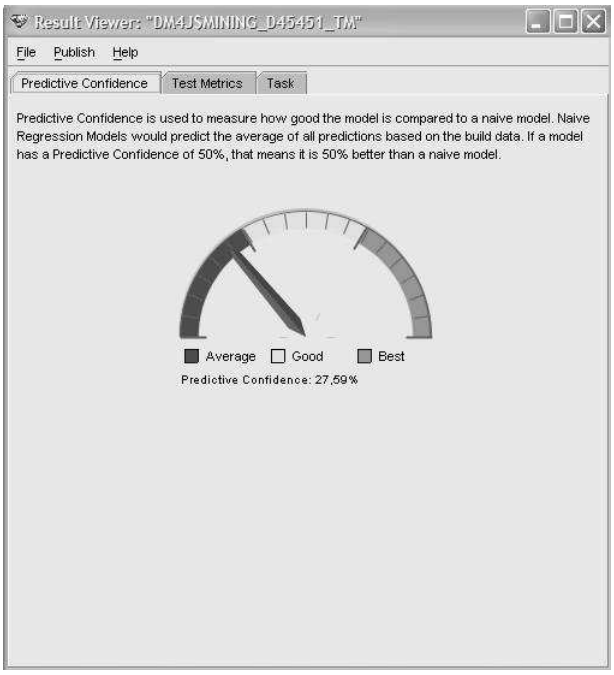

12. Kliknij przycisk OK. Upewnij się, że opcja Run upon finish jest włączona. Kliknij przycisk Zakończ.



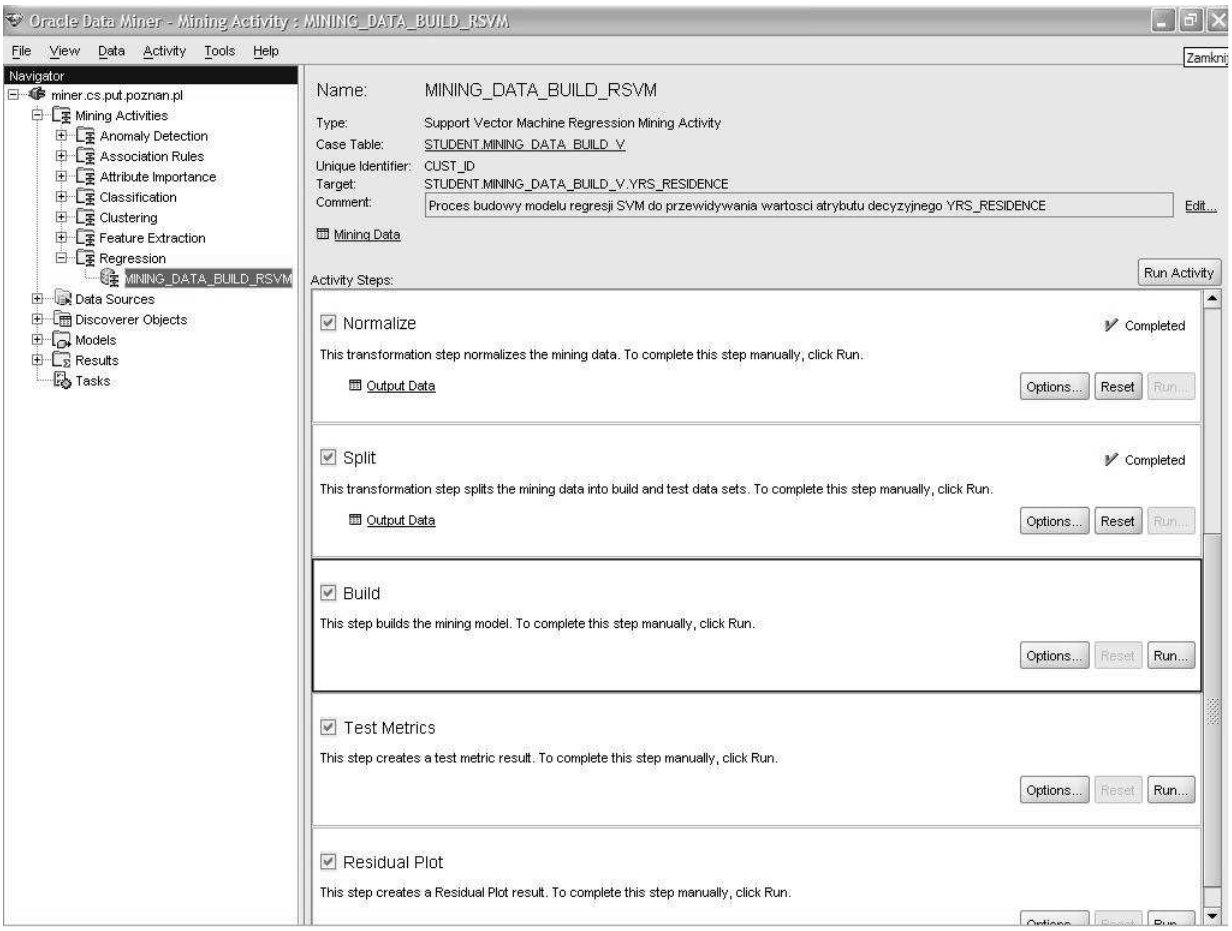
13. Kliknij na odnośnik Result w bloku Build. Współczynniki przy każdej wartości predyktorów definiują hiperpłaszczyznę najlepiej separującą instancje należące do klas decyzyjnych. Zauważ, że uzyskany wynik w praktyce nie poddaje się naturalnej interpretacji i stanowi rodzaj „czarnej skrzynki”.



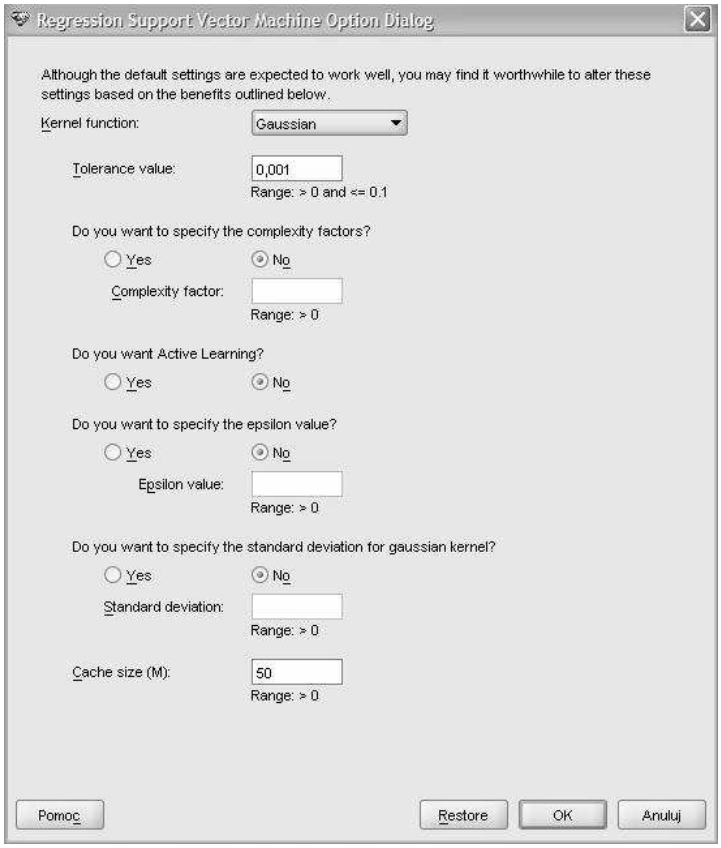

14. Zamknij okno z wynikami budowy klasyfikatora i powróć do głównego okna. Kliknij odnośnik Result w bloku Test Metrics. Na zakładce Predictive Confidence przedstawiona jest dokładność klasyfikatora liczona względem naiwnego klasyfikatora 0-R, który zawsze przewiduje najczęstszą wartość atrybutu decyzyjnego.




15. Powróć do głównego okna programu. Zaobserwuj zmianę jakości wygenerowanego klasyfikatora po korekcie parametrów algorytmu. Kliknij przycisk Reset w bloku Build (spowoduje to zresetowanie tego i wszystkich kolejnych kroków procesu odkrywania wiedzy).



16. Kliknij przycisk Options w bloku Build. Przejdź na zakładkę Algorithm Settings. Zmień rodzaj funkcji jądrowej na Gaussowską. Upewnij się, że opcja aktywnego uczenia się jest wyłączona.

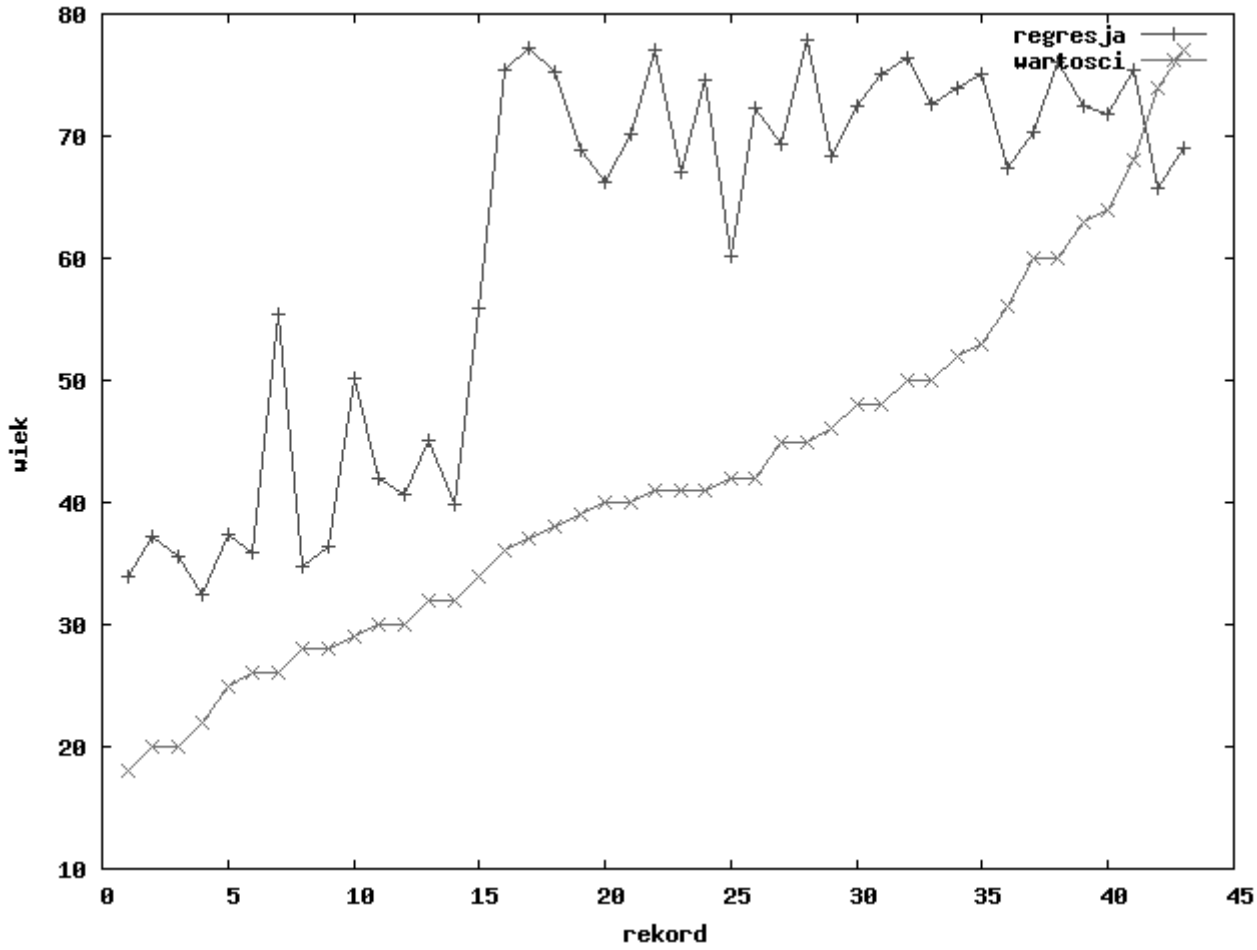



17. Powróć do głównego okna programu. Kliknij przycisk Run Activity (prawy górny okna). Po zakończeniu się procesu odkrywania wiedzy kliknij odnośnik Result w bloku Test Metrics. Czy nowy klasyfikator jest lepszy czy gorszy od poprzedniego?

## Ćwiczenie samodzielne

Wykorzystaj model stworzony za pomocą skryptu svm.reg.plsql do predykcji wieku klientów z Włoch. Jako źródło danych wykorzystaj tabelę `MINING_DATA_APPLY_V`. Pamiętaj o poddaniu danych źródłowych tym samym transformacjom, jakim były poddane dane, na podstawie których zbudowano model. Wykorzystując narzędzie GnuPlot dokonaj wizualizacji otrzymanych wyników.

Postaraj się uzyskać rezultat podobny do zamieszczonego poniżej:


regresja
wartosci
wiek
rekord
80
70
60
50
40
30
20
10
0
5
10
15
20
25
30
35
40
45