

## Laboratorium 12

### Odkrywanie osobliwości.

Odkrywanie osobliwości (ang. outliers) za pomocą algorytmu SVM zostanie w pierwszej części ćwiczenia przeprowadzone w środowisku SQL, a w drugiej części wykorzystamy narzędzie Oracle Data Miner.

1. Uruchom narzędzie iSQLPlus i połącz się z bazą danych.
2. Usuń obiekty pozostałe po poprzednim uruchomieniu algorytmu (konieczne w przypadku wielokrotnego uruchamiania procedury).

```
-- usuniecie obiektow pozostalych po poprzednich uruchomieniach
BEGIN
  EXECUTE IMMEDIATE 'DROP TABLE normalization';
  EXECUTE IMMEDIATE 'DROP VIEW v_prepared';
EXCEPTION
  WHEN OTHERS THEN NULL;
END;
/

-- usuniecie starej tabeli z ustawieniami algorytmu
BEGIN
  EXECUTE IMMEDIATE 'DROP TABLE settings';
EXCEPTION
  WHEN OTHERS THEN NULL;
END;
/
```

3. Dokonaj normalizacji atrybutów numerycznych AGE i YRS\_RESIDENCE. Poniższe zapytanie najpierw tworzy tabelę do przechowywania ustawień normalizacji, a następnie przeprowadza normalizację i tworzy perspektywę zawierającą dane po normalizacji.

```
BEGIN
-- utworzenie tabeli do przechowywania parametrow normalizacji
DBMS_DATA_MINING_TRANSFORM.CREATE_NORM_LIN (
  norm_table_name => 'normalization');

-- normalizacja za pomoca metody min-max
DBMS_DATA_MINING_TRANSFORM.INSERT_NORM_LIN_MINMAX (
  norm_table_name => 'normalization',
  data_table_name => 'mining_data_one_class_v',
  exclude_list   => DBMS_DATA_MINING_TRANSFORM.COLUMN_LIST('cust_id'));

-- utworzenie perspektywy pokazujacej znormalizowane dane
DBMS_DATA_MINING_TRANSFORM.XFORM_NORM_LIN (
  norm_table_name => 'normalization',
  data_table_name => 'mining_data_one_class_v',
  xform_view_name => 'v_prepared');
END;
/
```

4. Wyświetl fragment danych wejściowych po normalizacji.

```
SELECT *
FROM v_prepared
WHERE ROWNUM < 10;
/
```

Workspace

Enter SQL, PL/SQL and SQL\*Plus statements.

```
select cust_id, age, yrs_residence
from v_prepared
where rownum < 10
```

Execute Load Script Save Script Cancel

CUST_ID	AGE	YRS_RESIDENCE	
101504	.37037037		416666667
101505	.166666667		416666667
101510	.037037037		25
101520	.259259259		416666667
101522	.259259259		.333333333
101526	.277777778		.333333333
101530	.166666667		25
101535	.444444444		.333333333

5. Utwórz tabelę do przechowywania ustawień algorytmu i wypełnij ją parametrami algorytmu odkrywania osobliwości (nazwa algorytmu, typ funkcji jądrowej).

```
CREATE TABLE settings (
  setting_name VARCHAR2(30),
  setting_value VARCHAR2(30));

BEGIN
  INSERT INTO settings (setting_name, setting_value) VALUES
    (dbms_data_mining.algo_name,
 dbms_data_mining.algo_support_vector_machines);
  INSERT INTO settings (setting_name, setting_value) VALUES
    (dbms_data_mining.svms_kernel_function, dbms_data_mining.svms_linear);
  COMMIT;
END;
/
```

6. Usuń model z repozytorium (jeśli tworzyła(e)s model wcześniej)

```
BEGIN
  DBMS_DATA_MINING.DROP_MODEL('SVMO_Model');
EXCEPTION
  WHEN OTHERS THEN NULL;
END;
/
```

## 7. Utwórz model odkrywania osobliwości

```
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'SVMO_Model',
    mining_function     => dbms_data_mining.classification,
    data_table_name    => 'v_prepared',
    case_id_column_name => 'cust_id',
    target_column_name => NULL,
    settings_table_name => 'settings');
END;
/
```

## 8. Sprawdź sygnaturę modelu i zobacz, które atrybuty są wykorzystywane przy identyfikacji osobliwości.

```
SELECT attribute_name, attribute_type
FROM TABLE(DBMS_DATA_MINING.GET_MODEL_SIGNATURE('SVMO_Model'))
ORDER BY attribute_name;
```

## 9. Wyświetl szczegółowe informacje o modelu. W tym przypadku modelem jest unarny klasyfikator SVM a zawartością modelu są współczynniki określające kształt i położenie hiperpłaszczyzny separujące instancje należące do ogólnej klasy (1) od osobliwości.

```
WITH
mod_dtls AS (
  SELECT *
  FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_SVM('SVMO_Model'))
),
model_details AS (
SELECT D.class, A.attribute_name, A.attribute_value, A.coefficient
FROM mod_dtls D, TABLE(D.attribute_set) A
ORDER BY D.class, ABS(A.coefficient) DESC
)
SELECT class, attribute_name AS aname, attribute_value AS aval,
       coefficient AS coeff
FROM model_details;
```

Workspace

Enter SQL, PL/SQL and SQL\*Plus statements. Clear

```

WITH
mod_dtls AS (
  SELECT * FROM TABLE
(DBMS_DATA_MINING.GET_MODEL_DETAILS_SVM(SVMO_Model))
),
model_details AS (
  SELECT D.class, A.attribute_name, A.attribute_value, A.coefficient
  FROM mod_dtls D, TABLE(D.attribute_set) A
  ORDER BY D.class, ABS(A.coefficient) DESC
)

```

Execute Load Script Save Script Cancel

CLASS	ANAME	AVAL	COEFF
1	CUST_GENDER	M	1.43936882
1	COUNTRY_NAME	United States of America	1.41590442
1	CUST_MARITAL_STATUS	Married	1.23537045
1	CUST_GENDER	F	1.17233636
1			-1.0004971
1	YRS_RESIDENCE		990015697
1	HOUSEHOLD_SIZE	3	883978428
1	AGE		80516000

10. Zastosuj zbudowany model do znalezienia 10 klientów którzy najbardziej odróżniają się od całej reszty populacji.

```

SELECT cust_id
FROM (
  SELECT cust_id
  FROM v_prepared
  ORDER BY prediction_probability(SVMO_Model, 0 using *) DESC, 1)
WHERE rownum < 11;

```

11. Wyświetl przykładowe dane demograficzne o najbardziej typowych posiadaczach karty kredytowej, statystyki wyliczone na podstawie "typowych" klientów dają mniej zafałszowany obraz rzeczywistości, porównaj wynik z analogicznymi statystykami wyliczonymi na podstawie wszystkich klientów (zapytanie wyświetlające poniższe statystyki dla wszystkich klientów napisz samodzielnie).

```

SELECT a.cust_gender, ROUND(AVG(a.age)) AS age,
       ROUND(AVG(a.yrs_residence)) AS yrs_residence,
       COUNT(*) AS cnt
FROM mining_data_one_class_v a, v_prepared b
WHERE PREDICTION(SVMO_Model USING b.*) = 1
  AND a.cust_id=b.cust_id
GROUP BY a.cust_gender
ORDER BY a.cust_gender;

```

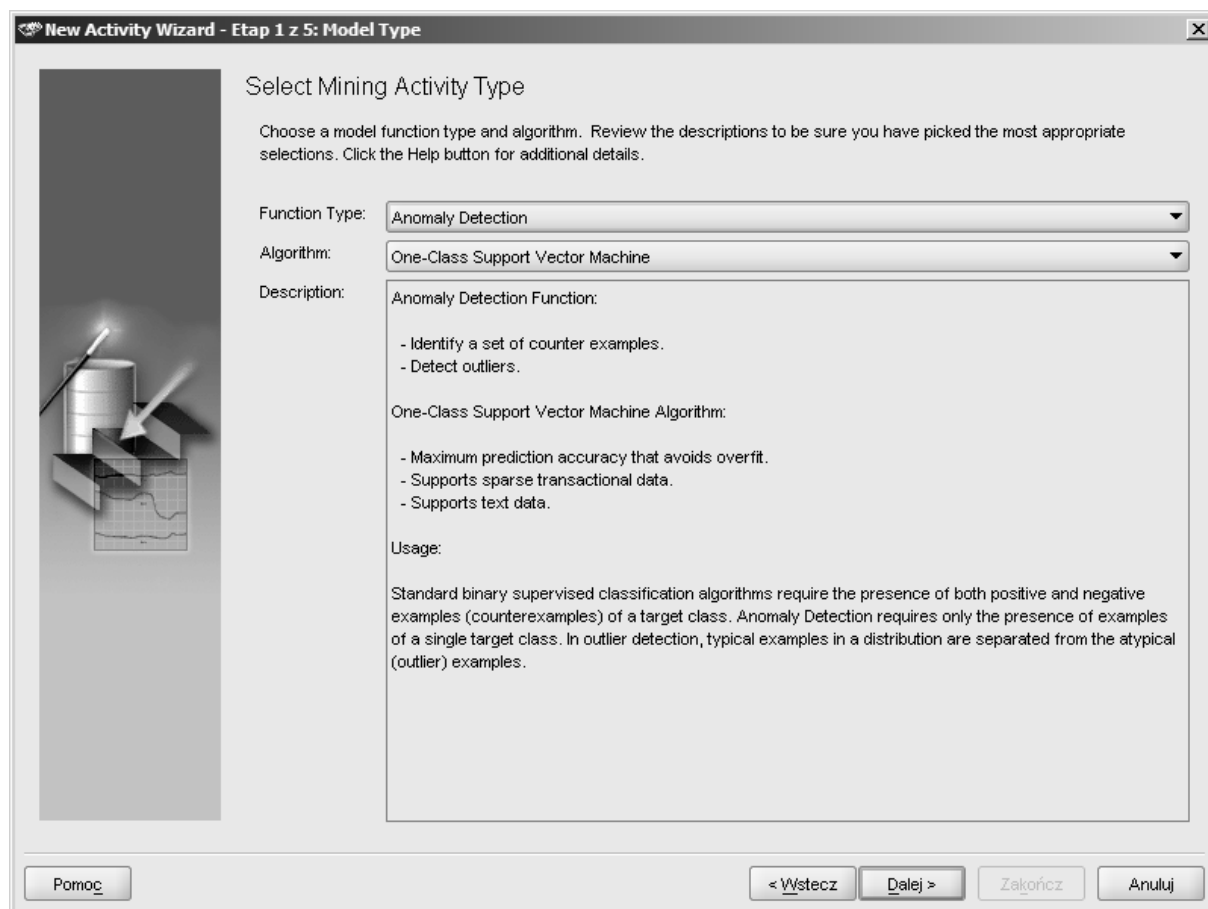
12. Wylicz prawdopodobieństwo, że dany klient jest “typowym” reprezentantem posiadaczy karty kredytowej.

```
WITH age_norm AS (  
SELECT shift, scale FROM normalization WHERE col = 'AGE'),  
yrs_residence_norm AS (  
SELECT shift, scale FROM normalization WHERE col = 'YRS_RESIDENCE')  
SELECT PREDICTION_PROBABILITY(SVMO_Model, 1 USING  
  (44 - a.shift)/a.scale AS age,  
  (6 - b.shift)/b.scale AS yrs_residence,  
  'Bach.' AS education,  
  'Married' AS cust_marital_status,  
  'Exec.' AS occupation,  
  'United States of America' AS country_name,  
  'M' AS cust_gender,  
  'L: 300,000 and above' AS cust_income_level,  
  '3' AS household_size  
  ) prob_typical  
FROM age_norm a, yrs_residence_norm b;
```

13. Uruchom narzędzie Oracle Data Mining i połącz się z bazą danych.

14. Z menu głównego wybierz **Activity**→**Build**. Na ekranie powitalnym kliknij przycisk **Dalej**>.

15. Z listy **Function Type** wybierz **Anomaly Detection**. Rozwiń listę **Algorithm** i wybierz z niej algorytm **One-Class Support Vector Machine**. Kliknij przycisk **Dalej**>.



**New Activity Wizard - Etap 1 z 5: Model Type**

Select Mining Activity Type

Choose a model function type and algorithm. Review the descriptions to be sure you have picked the most appropriate selections. Click the Help button for additional details.

Function Type: Anomaly Detection

Algorithm: One-Class Support Vector Machine

Description:

Anomaly Detection Function:

- Identify a set of counter examples.
- Detect outliers.

One-Class Support Vector Machine Algorithm:

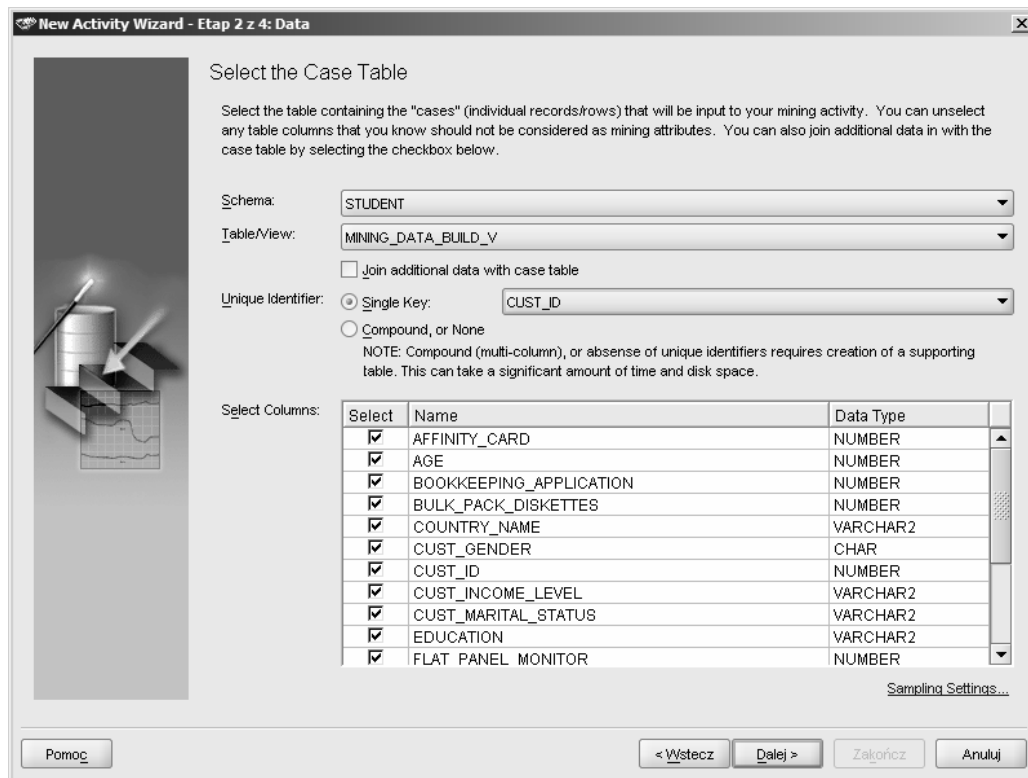
- Maximum prediction accuracy that avoids overfit.
- Supports sparse transactional data.
- Supports text data.

Usage:

Standard binary supervised classification algorithms require the presence of both positive and negative examples (counterexamples) of a target class. Anomaly Detection requires only the presence of examples of a single target class. In outlier detection, typical examples in a distribution are separated from the atypical (outlier) examples.

Pomoc < Wstecz Dalej > Zakończ Anuluj

16. Wskaż schemat STUDENT i tabelę MINING\_DATA\_BUILD\_V jako źródło danych do eksploracji. Jako klucz podstawowy wskaż atrybut CUST\_ID. Kliknij przycisk **Dalej>**.



Select the Case Table

Select the table containing the "cases" (individual records/rows) that will be input to your mining activity. You can unselect any table columns that you know should not be considered as mining attributes. You can also join additional data in with the case table by selecting the checkbox below.

Schema:

Table/View:

Join additional data with case table

Unique Identifier:  Single Key:   
 Compound, or None

NOTE: Compound (multi-column), or absence of unique identifiers requires creation of a supporting table. This can take a significant amount of time and disk space.

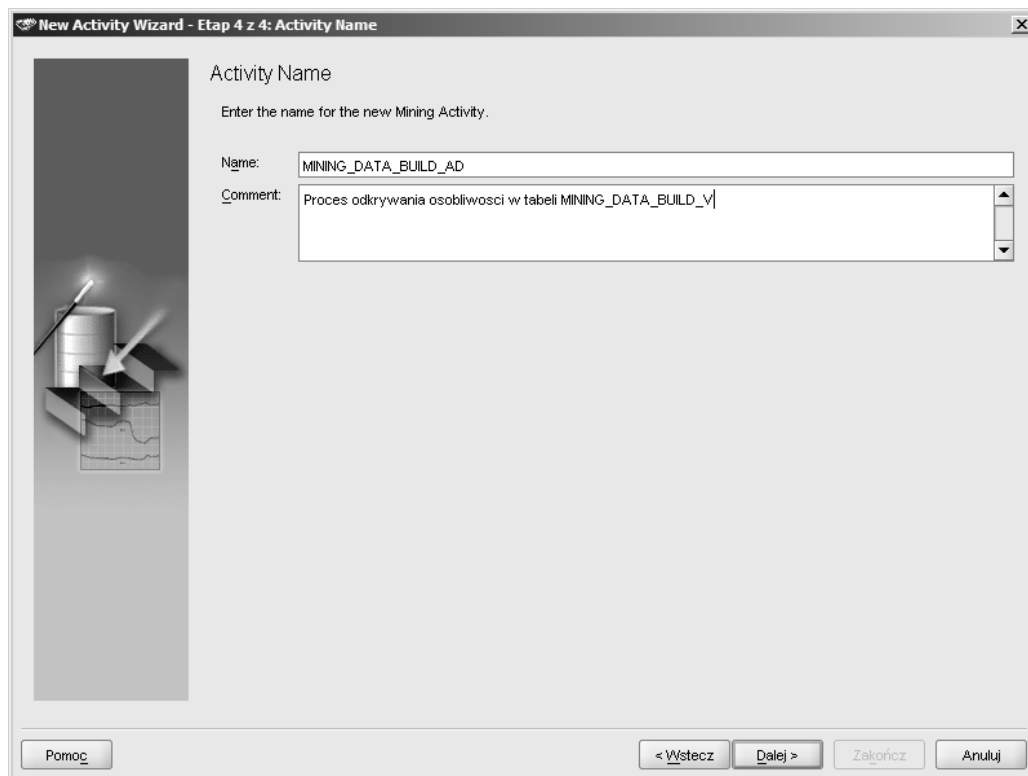
Select Columns:

Select	Name	Data Type
<input checked="" type="checkbox"/>	AFFINITY_CARD	NUMBER
<input checked="" type="checkbox"/>	AGE	NUMBER
<input checked="" type="checkbox"/>	BOOKKEEPING_APPLICATION	NUMBER
<input checked="" type="checkbox"/>	BULK_PACK_DISKETTES	NUMBER
<input checked="" type="checkbox"/>	COUNTRY_NAME	VARCHAR2
<input checked="" type="checkbox"/>	CUST_GENDER	CHAR
<input checked="" type="checkbox"/>	CUST_ID	NUMBER
<input checked="" type="checkbox"/>	CUST_INCOME_LEVEL	VARCHAR2
<input checked="" type="checkbox"/>	CUST_MARITAL_STATUS	VARCHAR2
<input checked="" type="checkbox"/>	EDUCATION	VARCHAR2
<input checked="" type="checkbox"/>	FLAT_PANEL_MONITOR	NUMBER

[Sampling Settings...](#)

Pomoc < Wstecz Dalej > Zakończ Anuluj

17. W kolejnym kroku podaj nazwę dla procesu eksploracji oraz krótki opis procesu eksploracji. Kliknij przycisk **Dalej>**.



Activity Name

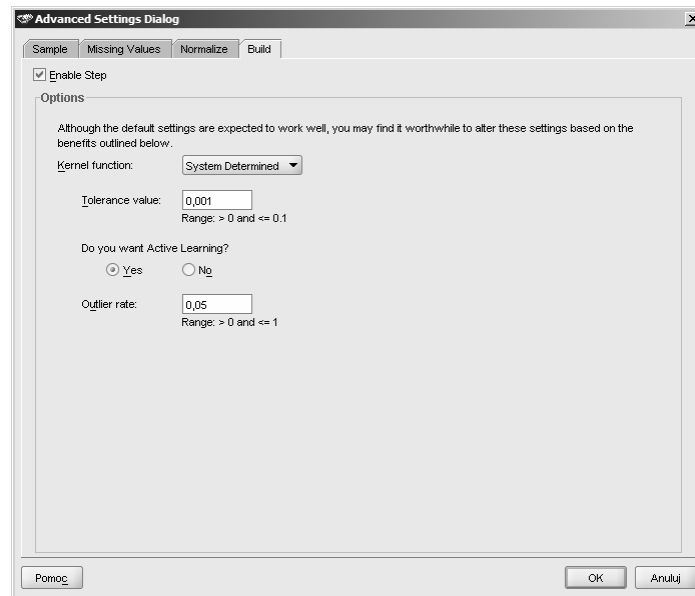
Enter the name for the new Mining Activity.

Name:

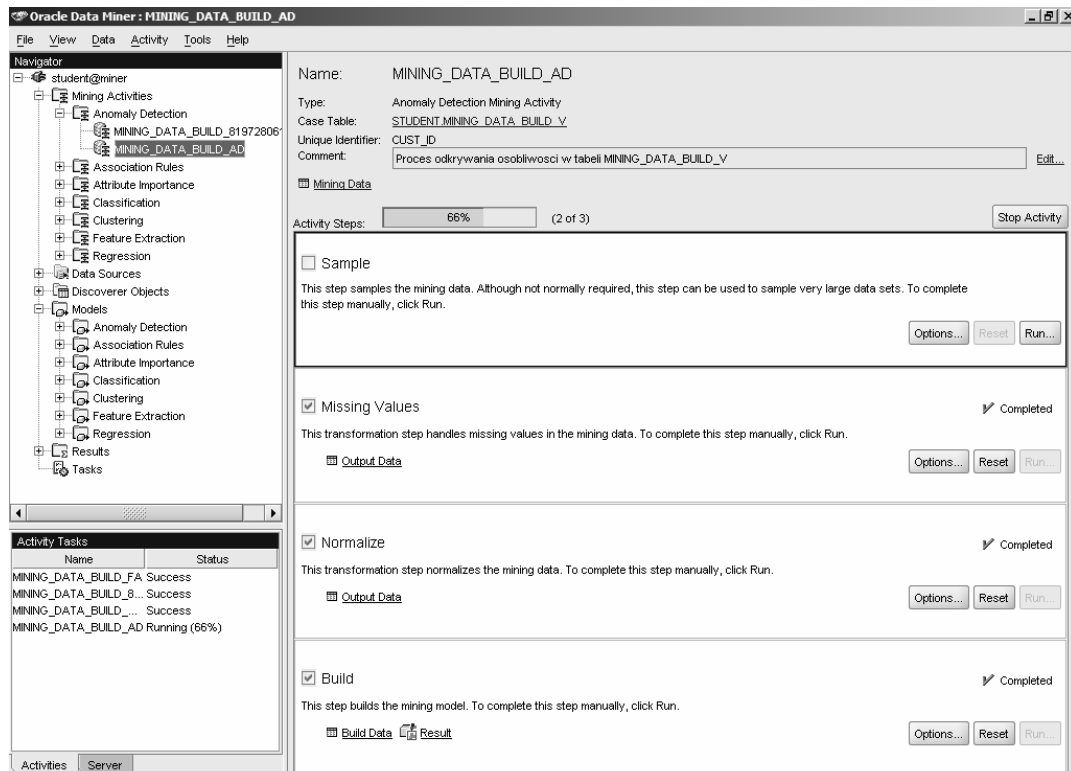
Comment:

Pomoc < Wstecz Dalej > Zakończ Anuluj

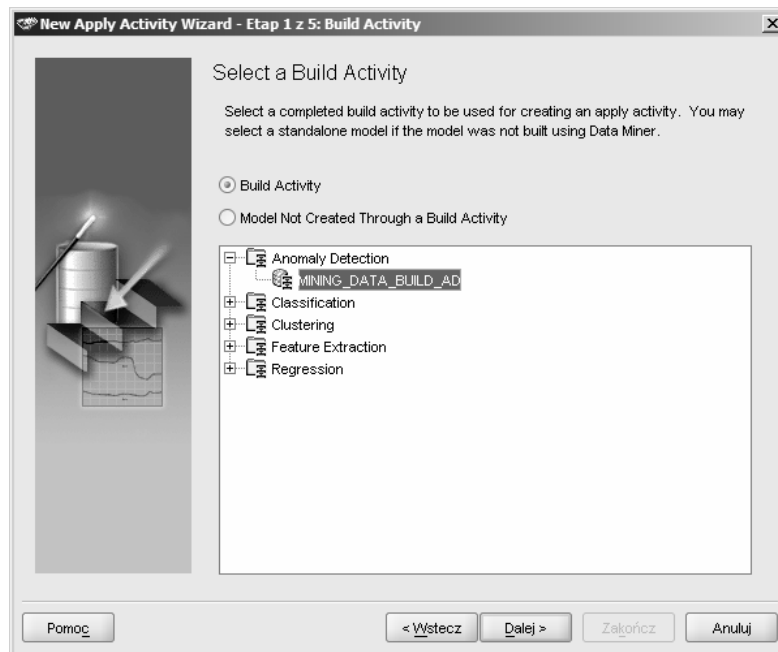
18. Kliknij przycisk **Advanced Settings**. Upewnij się, że na zakładce **Sample** opcja próbkowania jest wyłączona (pole wyboru **Enable Step** jest odznaczone). Przejdź na zakładkę **Missing Values** i upewnij się, że brakujące wartości będą automatycznie zamieniane na średnią (dla atrybutów numerycznych) lub wartość modalną (dla atrybutów kategoriycznych). Przejdź na zakładkę **Normalize** i upewnij się, że atrybuty numeryczne będą automatycznie normalizowane do przedziału 0-1. Przejdź na zakładkę **Build** i zmień przewidywany odsetek osobliwości na 5% (ustaw opcję **Outlier rate** na 0,05).



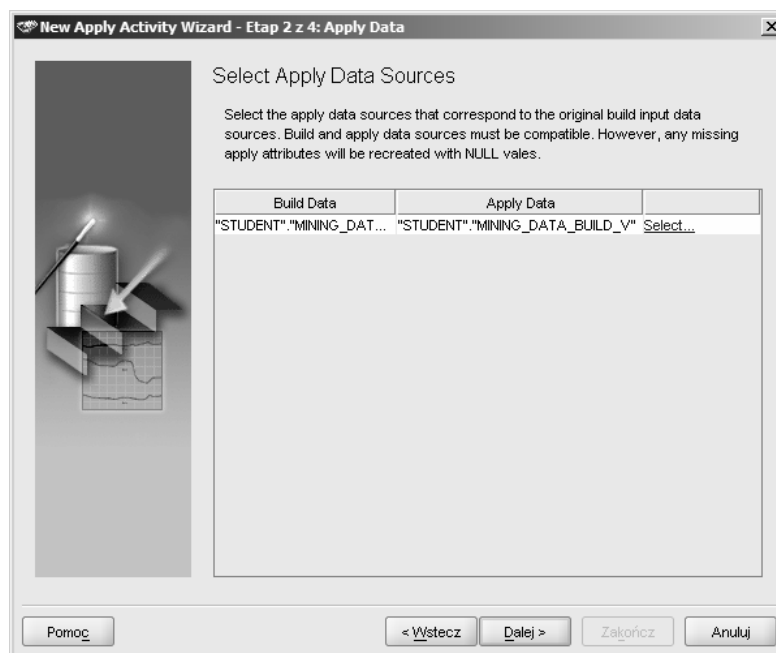
19. Kliknij przycisk **OK**. Upewnij się, że opcja **Run upon finish** jest włączona. Kliknij przycisk **Zakończ**.



20. Skonstruowany model nie jest dostępny do analizy i oglądu. W kolejnym kroku zastosujemy uzyskany model do identyfikacji osobliwości w wejściowym zbiorze danych. Z menu głównego wybierz **Activity**→**Apply**. Na ekranie powitalnym kliknij przycisk **Dalej**>. W pierwszym kroku asystenta wskaż na model, który ma zostać zastosowany. Upewnij się, że jest zaznaczone pole radiowe **Build Activity**. Rozwiń gałąź **Anomaly Detection** i zaznacz wcześniej wykonany proces. Kliknij przycisk **Dalej**>.

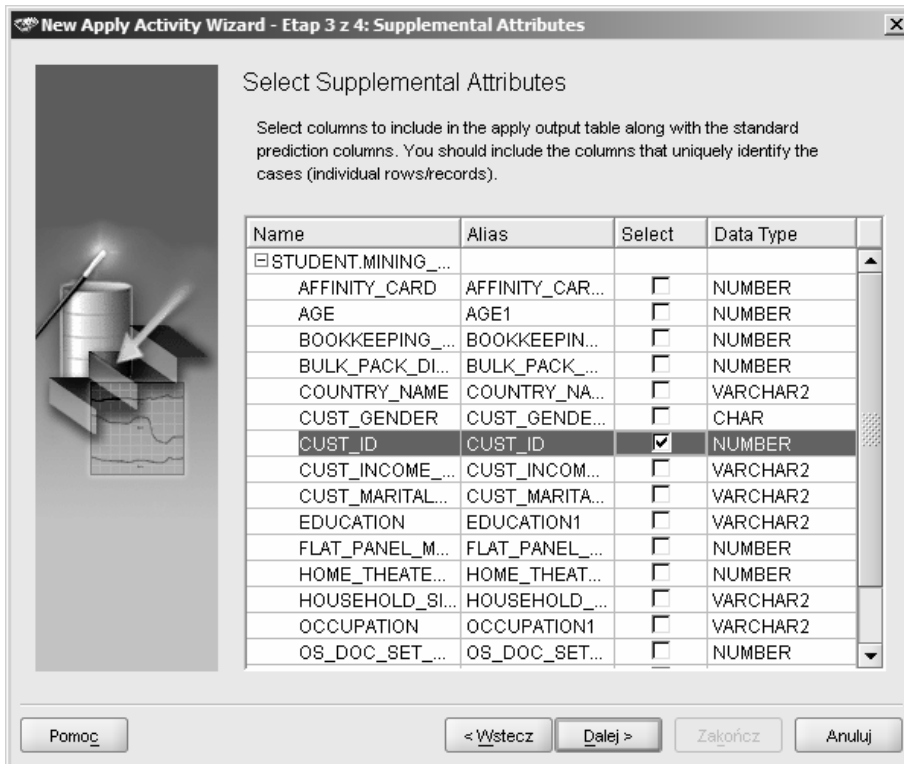


21. Kliknij na odnośnik **Select...** i wybierz schemat **STUDENT**, tabela **MINING\_DATA\_BUILD\_V**. Kliknij przycisk **Dalej**>.





22. Wskaż dodatkowe atrybuty, które powinny znaleźć się w tabeli wyjściowej. Z listy dostępnych atrybutów wybierz i zaznacz atrybut klucza podstawowego CUST\_ID. Kliknij przycisk **Dalej**.



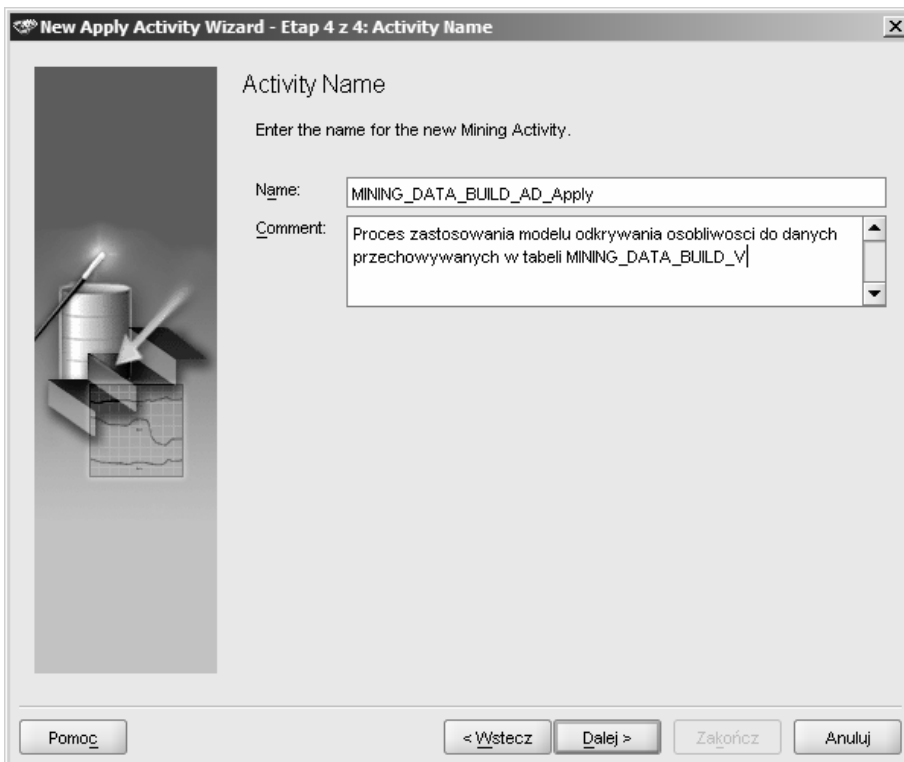
Select Supplemental Attributes

Select columns to include in the apply output table along with the standard prediction columns. You should include the columns that uniquely identify the cases (individual rows/records).

Name	Alias	Select	Data Type
STUDENT_MINING_...			
AFFINITY_CARD	AFFINITY_CAR...	<input type="checkbox"/>	NUMBER
AGE	AGE1	<input type="checkbox"/>	NUMBER
BOOKKEEPING_...	BOOKKEEPIN...	<input type="checkbox"/>	NUMBER
BULK_PACK_DI...	BULK_PACK_...	<input type="checkbox"/>	NUMBER
COUNTRY_NAME	COUNTRY_NA...	<input type="checkbox"/>	VARCHAR2
CUST_GENDER	CUST_GENDE...	<input type="checkbox"/>	CHAR
CUST_ID	CUST_ID	<input checked="" type="checkbox"/>	NUMBER
CUST_INCOME_...	CUST_INCOM...	<input type="checkbox"/>	VARCHAR2
CUST_MARITAL...	CUST_MARITA...	<input type="checkbox"/>	VARCHAR2
EDUCATION	EDUCATION1	<input type="checkbox"/>	VARCHAR2
FLAT_PANEL_M...	FLAT_PANEL_...	<input type="checkbox"/>	NUMBER
HOME_THEATE...	HOME_THEAT...	<input type="checkbox"/>	NUMBER
HOUSEHOLD_SI...	HOUSEHOLD_...	<input type="checkbox"/>	VARCHAR2
OCCUPATION	OCCUPATION1	<input type="checkbox"/>	VARCHAR2
OS_DOC_SET_...	OS_DOC_SET...	<input type="checkbox"/>	NUMBER

Pomoc < Wstecz Dalej > Zakończ Anuluj

23. W kolejnym kroku podaj nazwę dla procesu eksploracji oraz krótki opis procesu eksploracji. Kliknij przycisk **Dalej**.



Activity Name

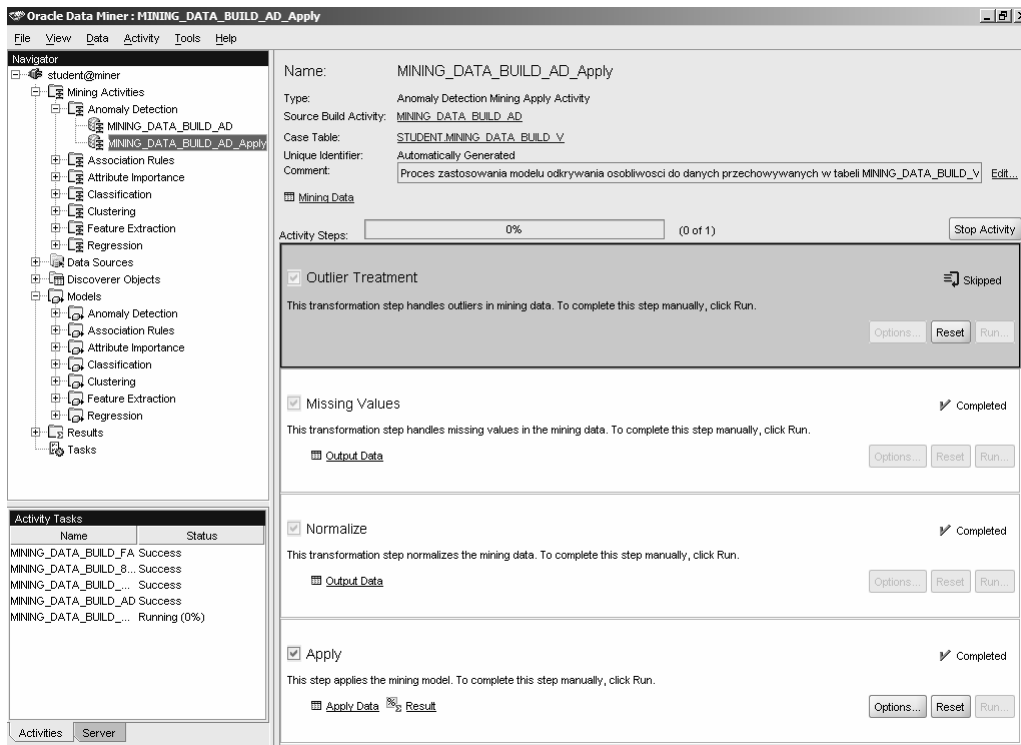
Enter the name for the new Mining Activity.

Name: MINING\_DATA\_BUILD\_AD\_Apply

Comment: Proces zastosowania modelu odkrywania osobliwosci do danych przechowywanych w tabeli MINING\_DATA\_BUILD\_V|

Pomoc < Wstecz Dalej > Zakończ Anuluj

## 24. Kliknij przycisk **Zakończ**.



25. Kliknij na odnośnik **Result** w bloku **Apply**. Osobliwościami są obiekty przypisane do klasy 0. Sprawdź, z jakim prawdopodobieństwem klasyfikator oznacza wybrane obiekty jako osobliwości. Wybierz jednego z klientów zaklasyfikowanego jako osobliwość i obejrzyj, za pomocą narzędzia iSQLPlus, całość informacji o wybranym przez siebie kliencie.

The screenshot shows the 'Result Viewer' window for the activity 'MINING\_DATA\_BUILD\_559055704\_A'. It displays a table with columns for DMR\$CASE\_ID, CUST\_ID, PREDICTION, and PROBABILITY. The table contains 23 rows of data.

DMR\$CASE_ID	CUST_ID	PREDICTION	PROBABILITY
101 501	101 501	1	0,5337
101 502	101 502	1	0,5232
101 503	101 503	1	0,5238
101 504	101 504	1	0,5271
101 505	101 505	1	0,5006
101 506	101 506	1	0,5355
101 507	101 507	1	0,5272
101 508	101 508	1	0,5216
101 509	101 509	1	0,5117
101 510	101 510	1	0,5068
101 511	101 511	1	0,5346
101 512	101 512	1	0,5272
101 513	101 513	1	0,5372
101 514	101 514	1	0,5376
101 515	101 515	1	0,5225
101 516	101 516	1	0,5383
101 517	101 517	1	0,5279
101 518	101 518	0	0,5056
101 519	101 519	1	0,5016
101 520	101 520	1	0,5207
101 521	101 521	1	0,5369
101 522	101 522	1	0,5161
101 523	101 523	1	0,5079