**Politechnika Świętokrzyska**

**Sprawozdanie z projektu**

Przedmiot: Hurtownie i eksploracja danych

Autor: Bartosz Dygas

Grupa: 1ID21A

1. **Opis słowny**

Hurtownia jest zbudowana w oparciu o sieć zakładów samochodowych. Dane w hurtowni są mapowane w strukturę gwiazdy. Zakłady obsługują w poszczególnych warsztatach, na poszczególnych halach i podnośnikach samochody należące do klientów. Do każdej naprawy wyznaczony jest jeden pracownik, który tą naprawą nadzoruje i po każdej naprawie generowana jest faktura. Naprawa jest też związana z awarią lub zestawem awarii, której uległ samochód i do której trzeba użyć konkretnych zestawów narzędzi i części. Migracja danych z bazy danych do hurtowni danych następuje raz dziennie.

Do tabel warsztatów, hal, podnosników – rekordy są dodawane bardzo rzadko. Następuje to zwykle kilka razy do roku gdy wybudowany zostanie nowy budynek firmy. Dodanie rekordów częściej następuje dla tabeli pracowników, narzędzi i części, natomiast dla tabel pojazdów, faktur, klientów rekordy są dodawane praktycznie codziennie.

Dla hurtowni przygotowanych zostało kilkanaście zapytań (w języku SQL), które pomogą w analizie danych zawartych w hurtowni. Zostały również zbudowane 3 modele eksploracji danych – klasyfikacja, asocjacja i grupowanie. Dzięki tym modelom można będzie pogrupować naprawy, zbadać jakie awarie najczęsciej następują wspólnie w danej marce samochodów oraz przewidzieć płeć klienta za pomocą narzędzia klasyfikacji.

1. **Projekt hurtowni danych**

Na tabele wymiarów składają się:

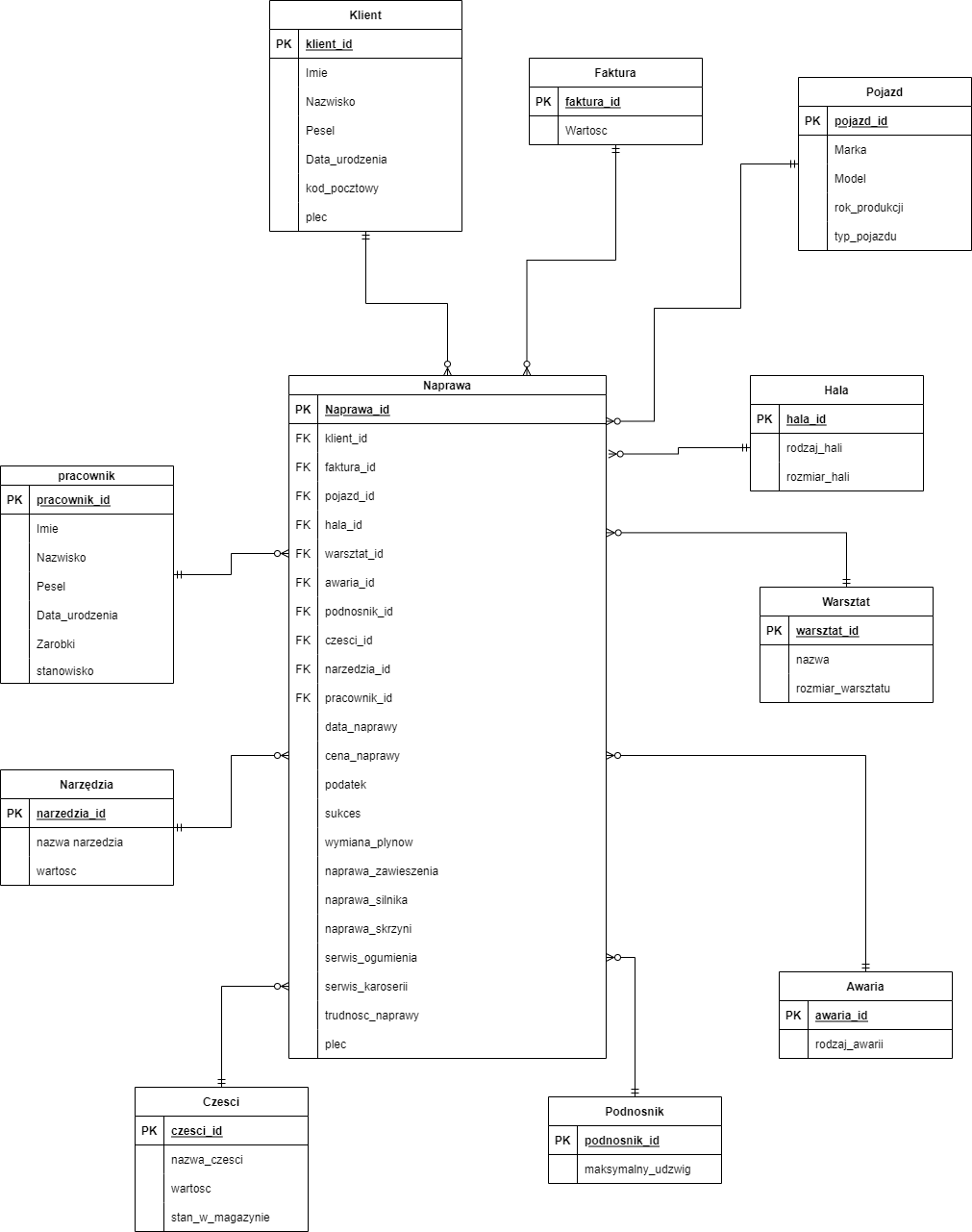
* Awaria - tabela zawierająca wszystkie spotkane rodzaje awarii samochodów
* Podnosnik – tabela zawiera wszystkie podnosniki uzywane w firmie
* Czesci – tabela zawiera zestawy czesci
* Narzedzia – tabela zawiera zestawy narzedzi
* Pracownik – wszyscy pracownicy
* Klient – wszyscy klienci
* Faktura – wszystkie faktury
* Pojazdów – wszystkie pojazdy z rozróżnieniem na markę, model, typ, rok produkcji
* Hala – rodzaje hal w warsztatach
* Warsztat – wszystkie warsztaty należące do firmy

Tabelą faktów jest tabela napraw. Zawiera ona klucze obce, za pomocą których łączy się z tabelami wymiarów oraz miary, które pozwalają na dokładniejszą eksplorację danych.

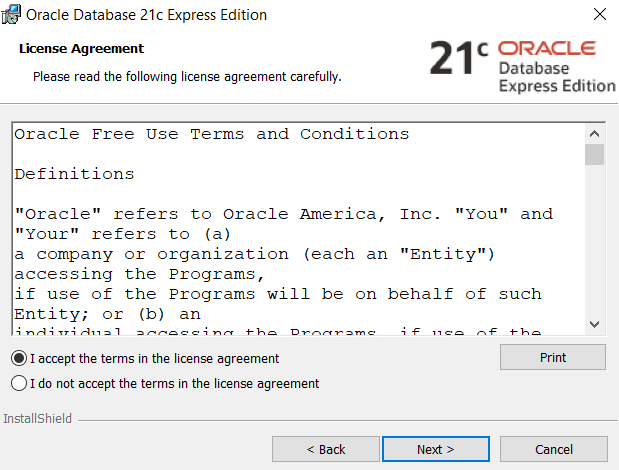
Miary tabeli faktów:

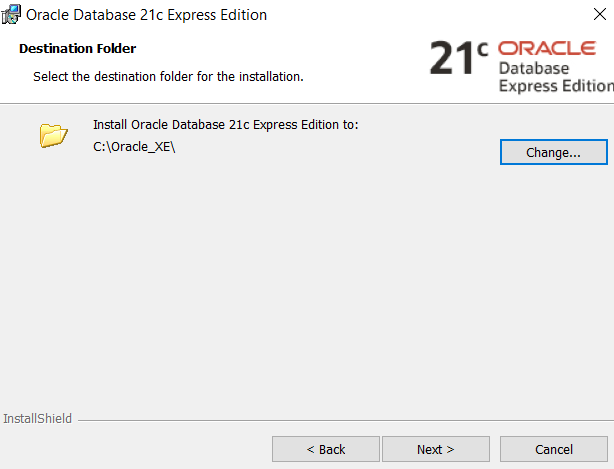
* Data\_naprawy – data w formacie rok/miesiąc/dzień
* Cena\_naprawy – całkowita cena jaką klient musi zapłacić za naprawę
* Podatek – podatek jaką zakład musi zapłacić za naprawę
* Sukces – przybiera wartości TAK/NIE i mówi o tym, czy naprawa się udała
* Wymiana\_płynów – przybiera wartości TAK/NIE i mówi o tym, czy podczas naprawy zostały wymienione płyny
* Naprawa\_zawieszenia – przybiera wartości TAK/NIE i mówi o tym, czy podczas naprawy było naprawiane zawieszenie
* Naprawa\_silnika – przybiera wartości TAK/NIE i mówi o tym, czy podczas naprawy był naprawiany silnik
* Naprawa\_skrzyni – przybiera wartości TAK/NIE i mówi o tym, czy podczas naprawy była naprawiana skrzynia
* Serwis\_ogumienia – przybiera wartości TAK/NIE i mówi o tym, czy podczas naprawy był dokonywany serwis opon
* Serwis\_karoserii – przybiera wartości TAK/NIE i mówi o tym, czy podczas naprawy były dokonywane naprawy blacharskie
* Trudnosc\_naprawy – przybiera wartości BARDZO\_LATWY/LATWY/PRZECIETNY/TRUDNY/BARDZO\_TRUDNY i mówi o tym, jak trudna i czasochłonna była naprawa według pracowników
* Plec – plec klienta

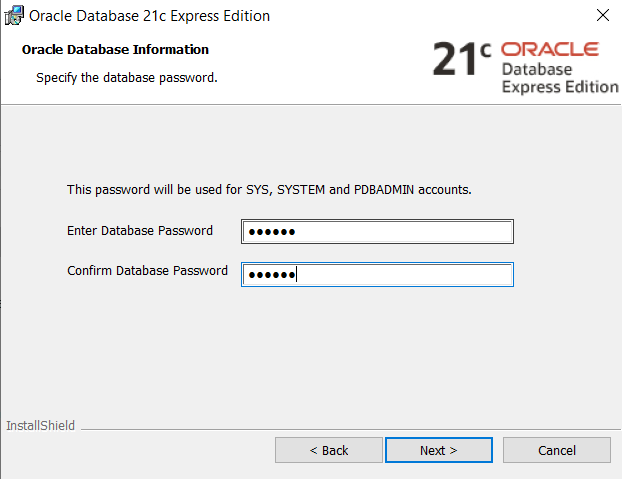
Schemat hurtowni danych

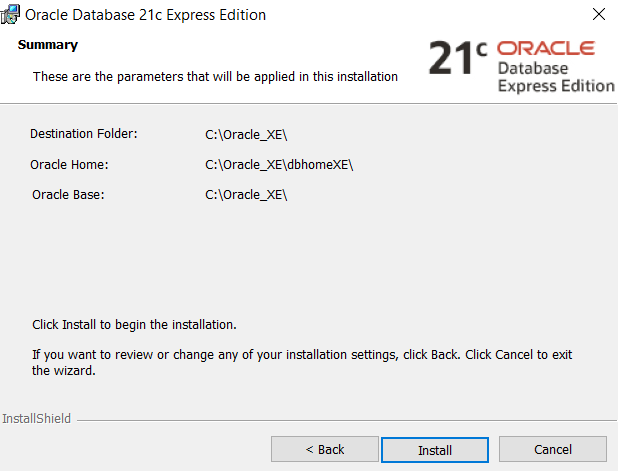


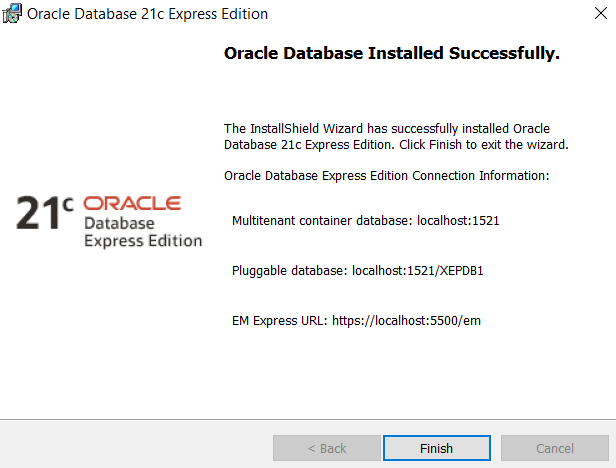
1. **Instalacja środowiska pracy**







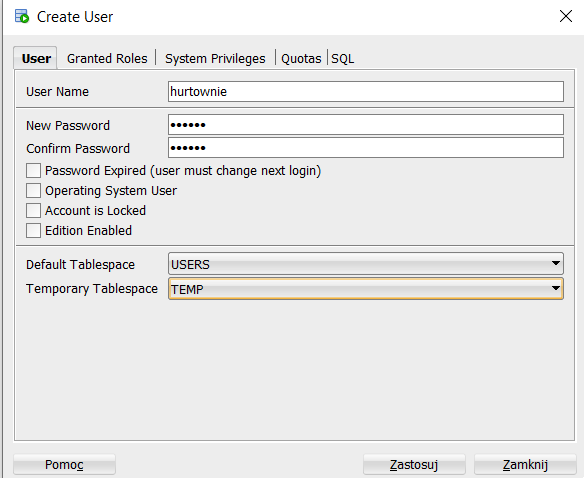




Po instalacji programu Oracle Database 21c Express Edition można skonfigurować Oracle developer. Pierwszym krokiem jest połączenie się z bazą danych, z kontem administratora:

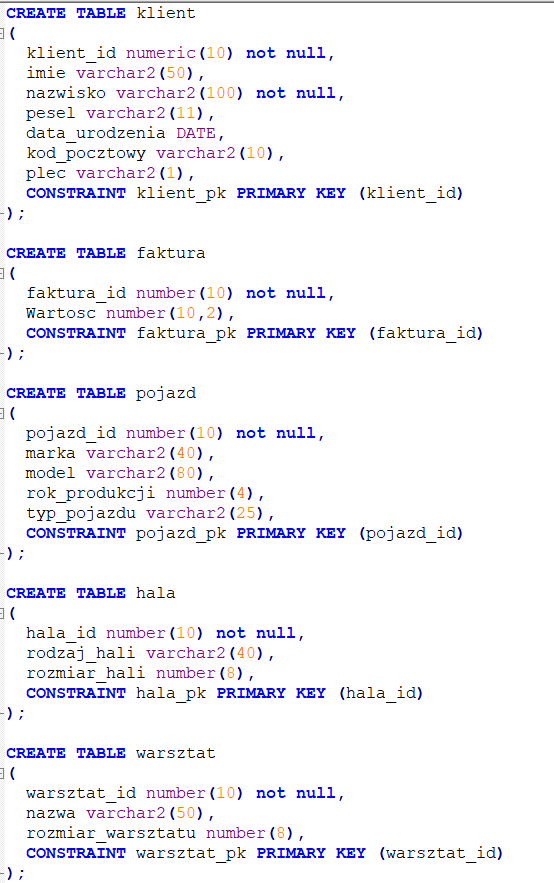


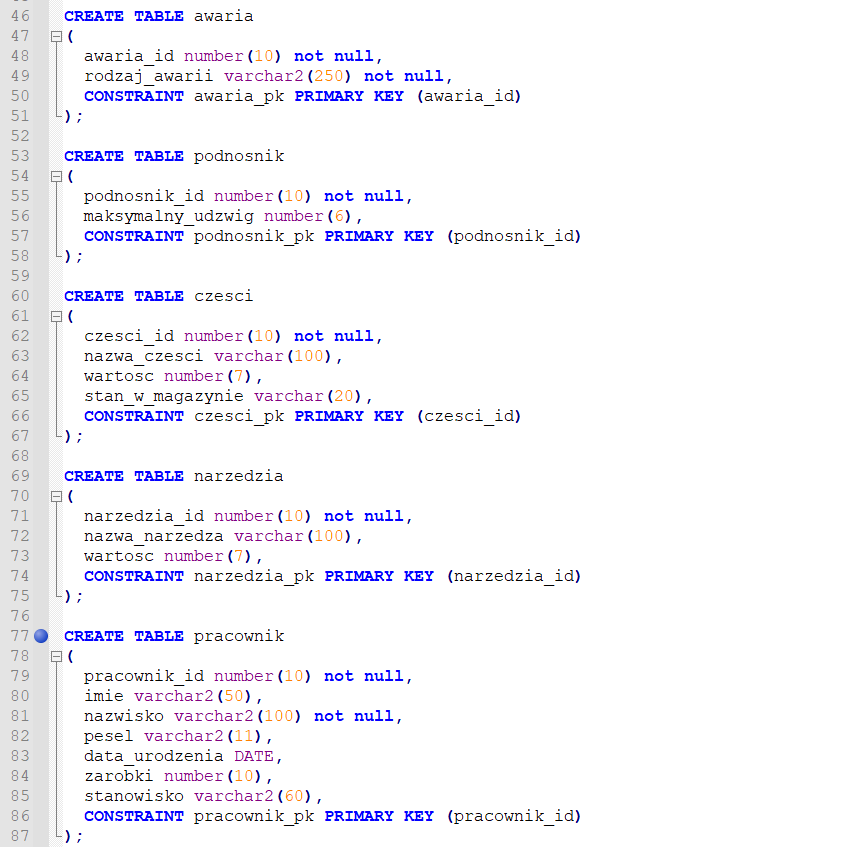
Następnie możliwe jest stworzenie nowego użytkownika, dla którego stworzymy hurtownię.



1. **Zasilenie hurtowni danych**

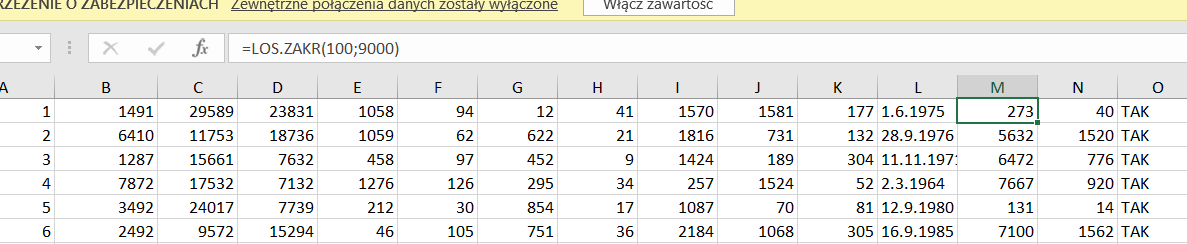
Aby zasilić hurtownię danych trzeba najpierw przygotować odpowiednie tabele.

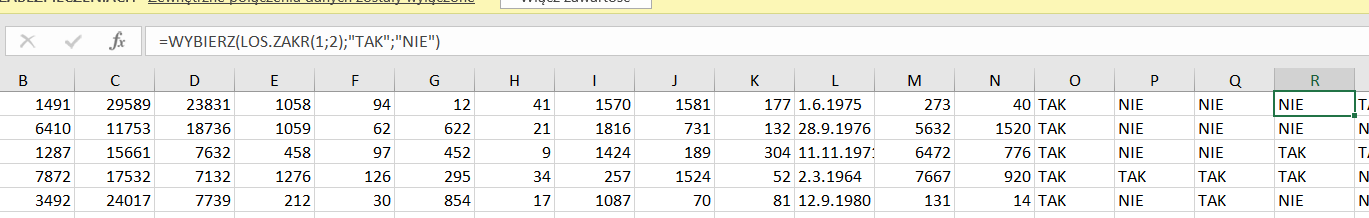






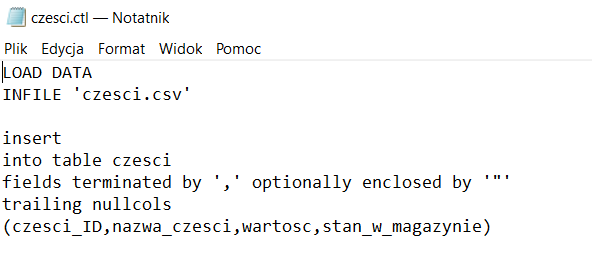
Kolejnym krokiem jest wygenerowanie odpowiednich danych. Dane zostały wygenerowane w programie Microsoft Excel za pomocą funkcji LOS.ZAKR oraz WYBIERZ.





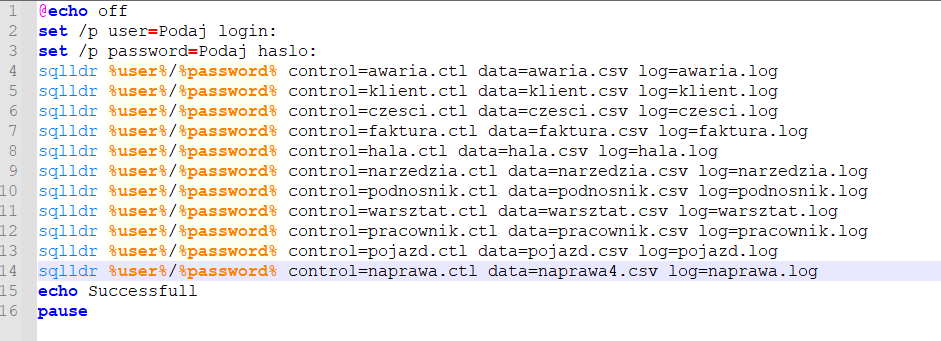
Do załadowania danych do tabel użyte zostały skrypty składające się z plików .ctl potrzebny do zlokalizowania danych oraz bazy, do której dane mają być wgrane oraz pliku .bat wgrywającym wszystkie dane do tabel.

Przykładowy plik .ctl:



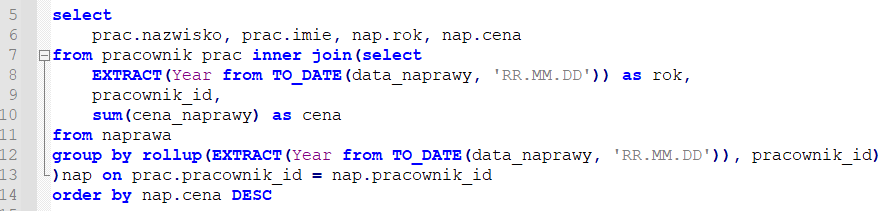
W pliku tym zawarta jest nazwa tabeli, z której wgrywać będziemy dane, tabela i jej kolumny w hurtowni danych, do których wgrywane będą dane oraz separator kolumn w pliku .csv.

Plik .bat:



Plik .bat wymaga podania loginu i hasła użytkownika. Za pomocą narzędzia sql loader pobiera dane z pliku .ctl oraz dane z pliku .csv. Dokonuje próby ładowania danych i wyniki umieszcza w pliku .log.

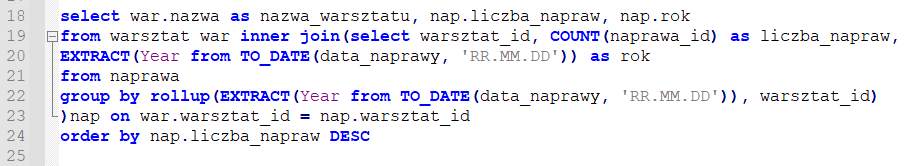
1. **Zapytania do hurtowni danych**
   1. Zapytania Rollup
      1. Pierwsze zapytanie typu Rollup sprawdza ile pieniędzy dla firmy wygenerowali poszczególni pracownicy z podziałem na lata. Dzięki temu można wynagrodzić najlepszych pracowników premiami i awansami.



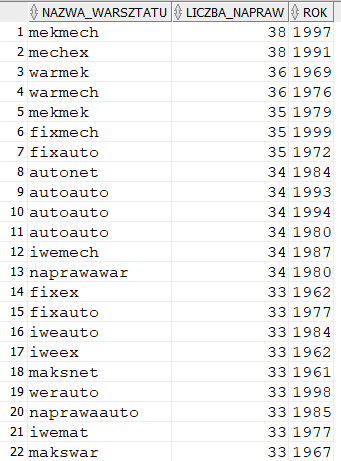
Wynik:

****

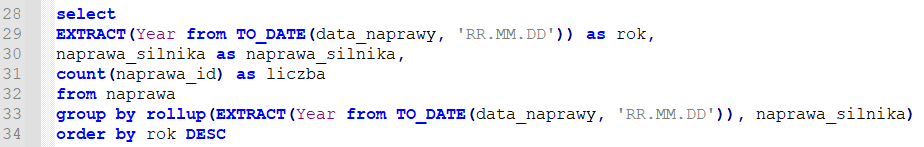
5.1.2. Zapytanie sprawdza ile napraw dokonały poszczególne warsztaty z podziałem na lata. Dzięki temu można na przykład zainwestować w zakłady, w których dokonuje się większej liczby napraw.



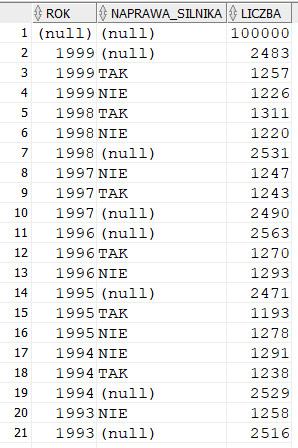
Wynik:



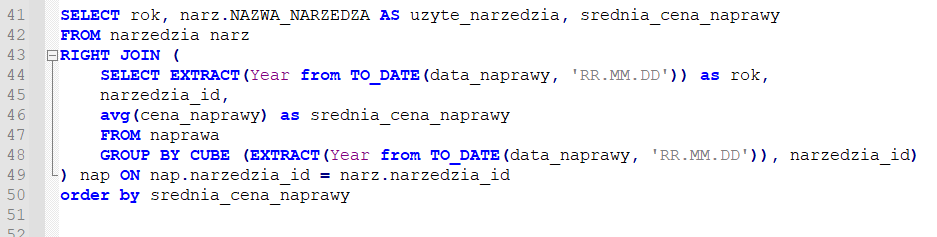
5.1.3. Ostatnie zapytanie Rollup sprawdza ile razy wykonana została naprawa silnika w poszczególnych latach. Pozwala to oszacować awaryjność silników w poszczególnych latach i na tej podstawie można przewidzieć prawdopodobne przyszłe rodzaje usterek i napraw.



Wynik:



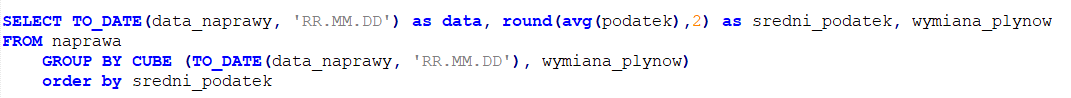
* 1. Zapytania Cube
     1. Zapytanie oblicza średnią cenę naprawy w zależności od zestawu użytych narzędzi do tej naprawy oraz w zależności od roku naprawy. Pozwala to na oszacowanie jakich narzędzi najlepiej używać w celu zminimalizowania kosztów naprawy.



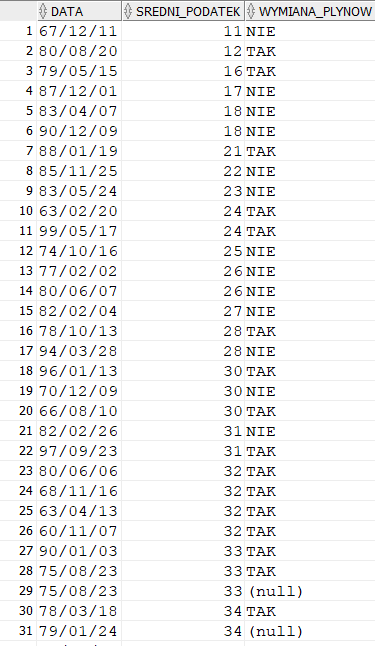
Wynik:



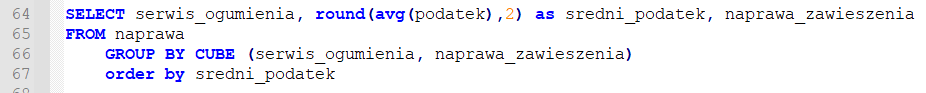
* + 1. Następne zapytanie pozwala na sprawdzenie zależności średniego zapłaconego podatku od wymiany płynów w konkretnym dniu.



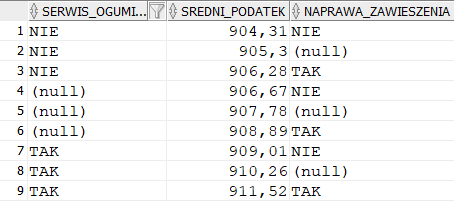
Wynik:



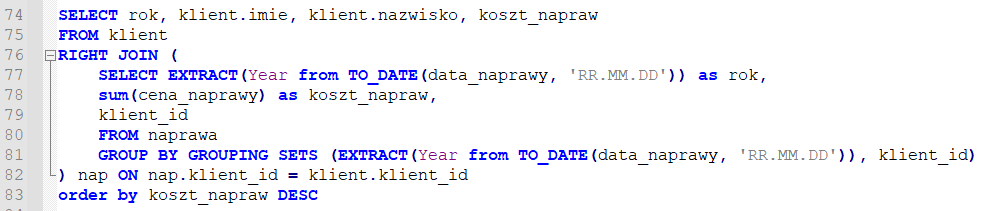
* + 1. Zapytanie pozwala sprawdzić zależność między naprawą zawieszenia, serwisem opon i średnim podatkiem. Dzięki temu firma może w przyszłości skupić się na serwisach, przez które może zapłacić mniejszy podatek



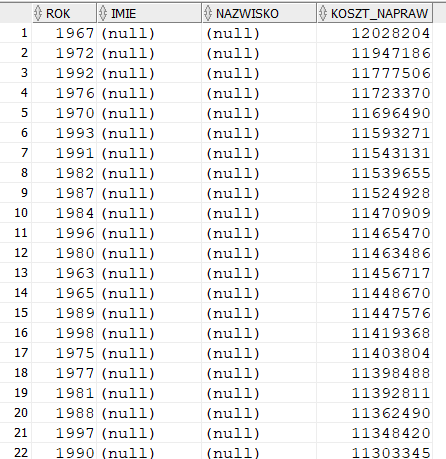
Wynik:



* 1. Grouping sets
     1. Zapytanie pozwala sprawdzić wydatki na naprawy poszczególnych klientów oraz w poszczególnych latach. Pomoże to zakładowi określi klientów którzy przynoszą największe zyski i zorientować się, w których latach zyski były najlepsze, a w których gorsze.



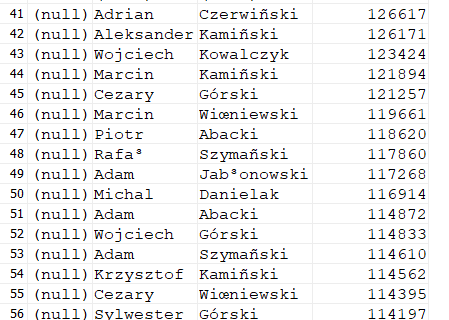
Wynik:



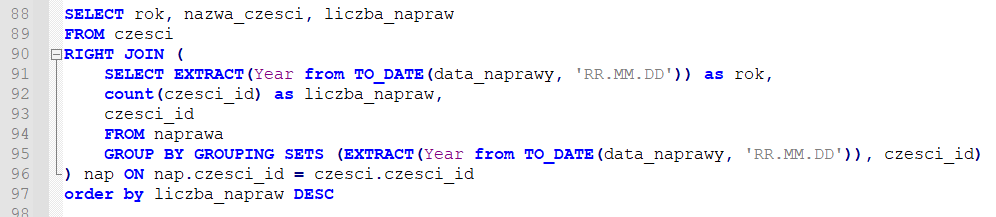
.

.

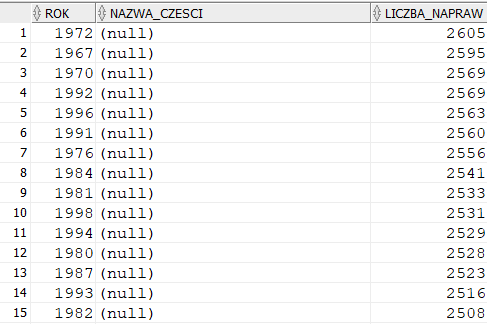
.



5.3.2 Liczba napraw w poszczególnych latach oraz liczba napraw z wykorzystaniem poszczególnych części. Pomoże to m.in. oszacować jakie części wykorzystywane są najczęściej i w jakie warto się zaopatrywać.



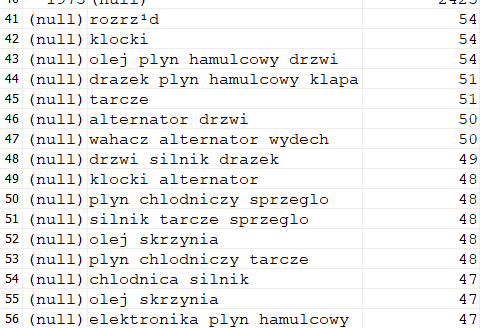
Wynik:



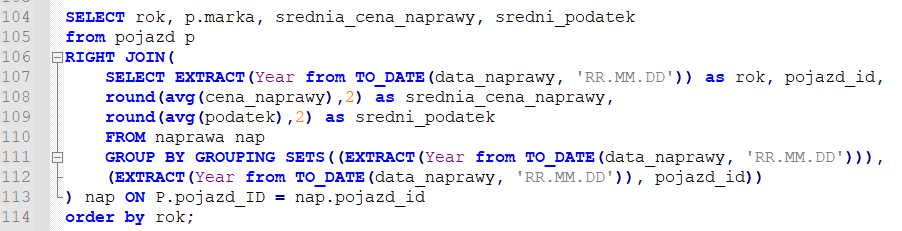
.

.

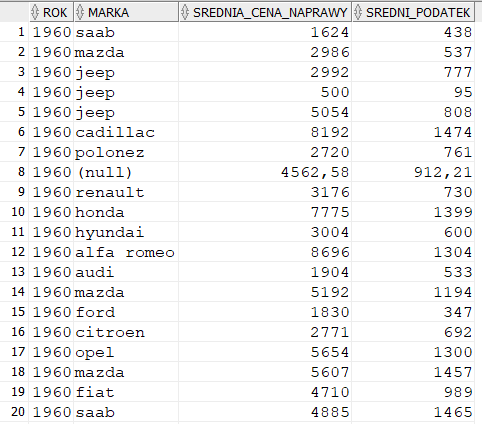
.



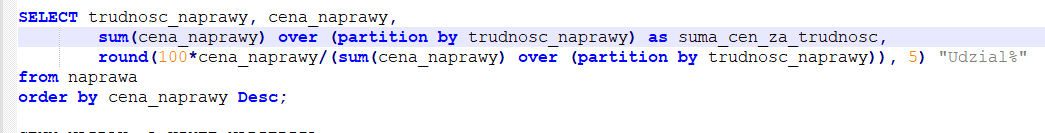
5.3.3 Zapytanie oblicza średnią cenę naprawy oraz średni podatek dla poszczególnych marek samochodów w poszczególnych latach. Pomaga to w oszacowaniu najbardziej dochodowych marek.



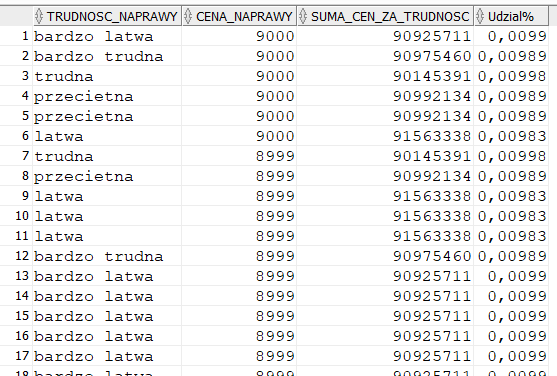
Wynik:



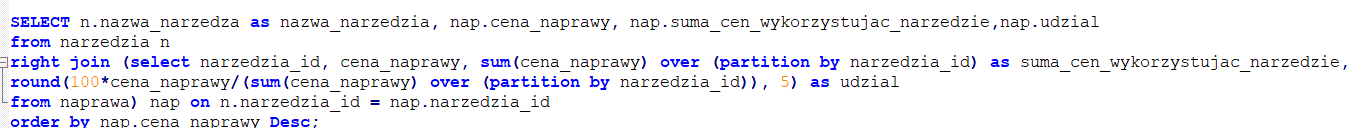
* 1. Partycje obliczeniowe
     1. Zapytanie sprawdza zależność trudności naprawy od jej ceny, wypisuje sumę cen za daną trudność oraz oblicza udział w tej sumie dla poszczególnego rekordu. Pomaga to w oszacowaniu czy trudniejsze naprawy są bardziej opłacalne, czy może warto z nich zrezygnować.



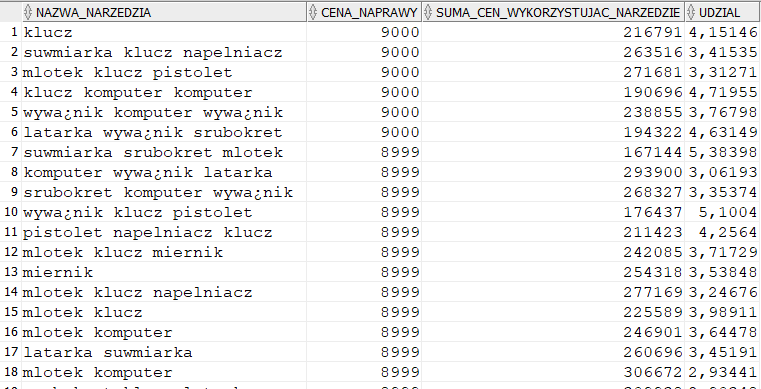
Wynik:



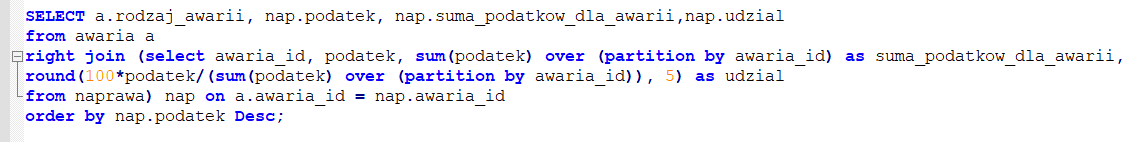
* + 1. Następne zapytanie porównuje cenę naprawy z wykorzystaniem różnych zestawów narzędzi. Dzięki temu możemy zobaczyć, które narzędzia używane są w naprawach przynoszących największy zysk.



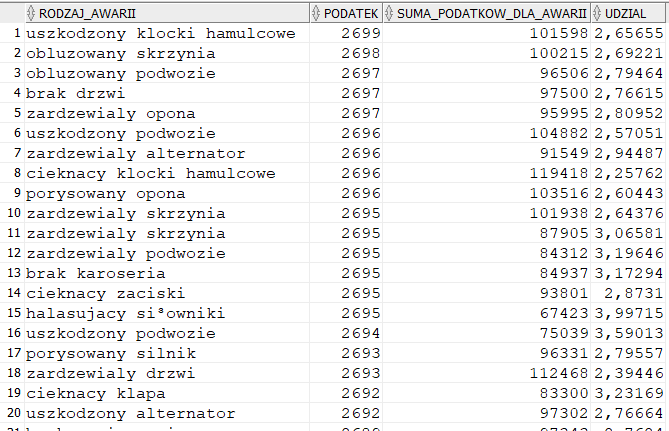
Wynik:

****

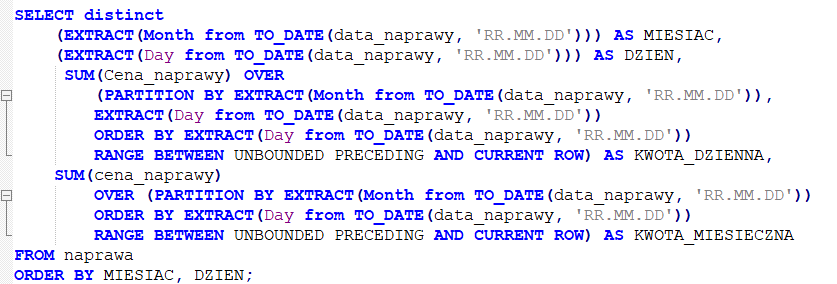
* + 1. Zapytanie pokazuje zależność między rodzajem awarii, a zapłaconym podatkiem. Dzięki temu można zrezygnować z napraw, które generują największe podatki dla zakładu.



Wynik:

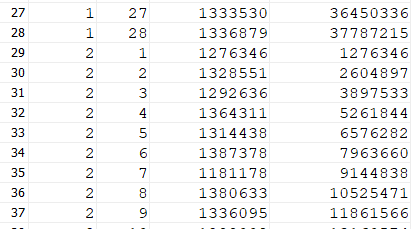


* 1. Okna ruchome
     1. Zapytanie pozwala na oszacowanie, w który dzień i miesiąc zakłady przynoszą największe zyski. Dzięki temu można lepiej wyznaczyć urlopy pracownikom, przestoje warsztatów, czy inne przerwy w pracy.

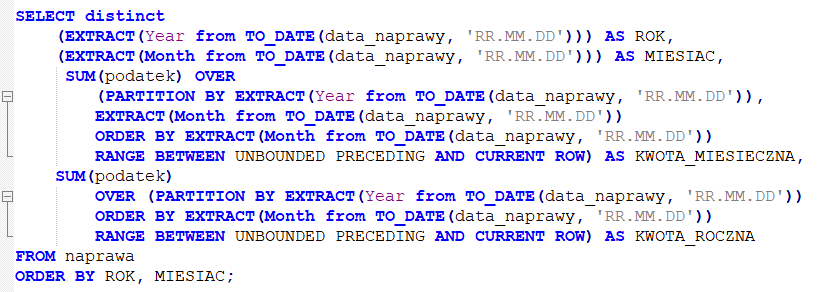


Wynik:

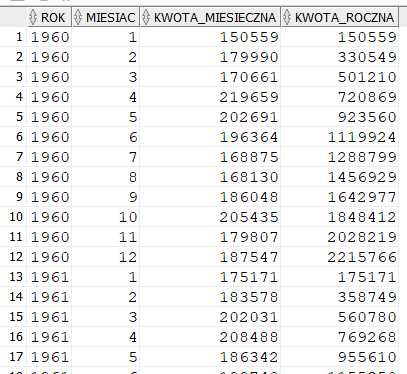




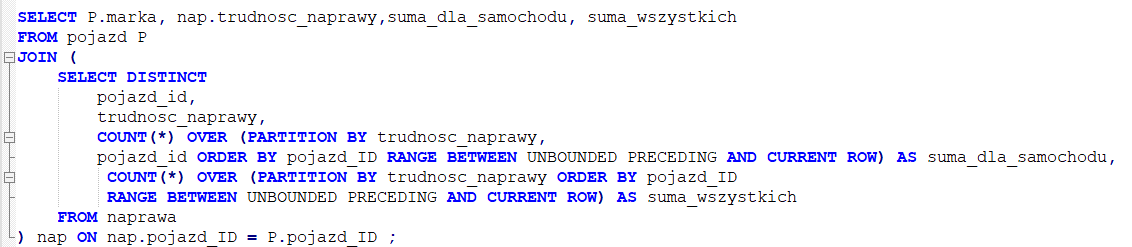
* + 1. Następne pytanie pozwala na określenie w którym miesiącu i roku płacone są najwyższe, a w którym najniższe podatki. To zapytanie może pomóc w decyzji, kiedy rozliczać się z podatków.



Wynik:



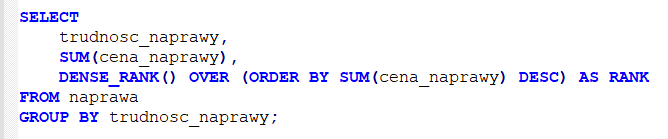
* + 1. Następne zapytanie pozwala na zliczenie trudnosci napraw dla poszczegolnych marek samochodow oraz ogólnym zliczeniu występujących trudności napraw. Pomoże to zorientować się w tym, jak trudne naprawy są obsługiwane najczęściej, oraz które samochody są najtrudniejsze, a które najłatwiejsze w naprawach.



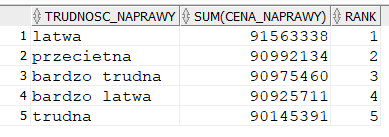
Wynik:



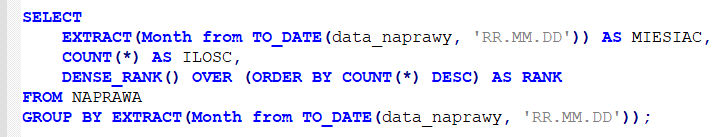
* 1. Zapytania rankingowe
     1. Pierwsze zapytanie rankingowe pomoże ustalić jakie rodzaje napraw ze względu na ich trudność przynoszą największe zyski



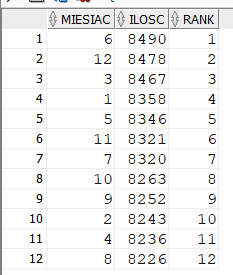
Wynik:



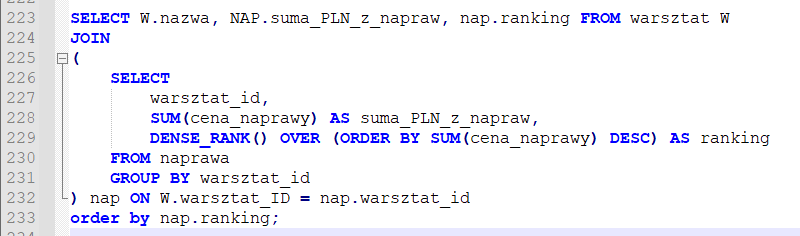
* + 1. Drugie zapytanie pozwala ustalić, w którym miesiącu występuje największa liczba napraw.



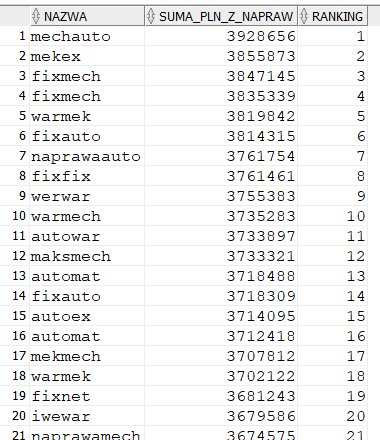
Wynik:



* + 1. Ostatnie zapytanie rankingowe pozwala ustalić, który warsztat generuje największe zyski z napraw



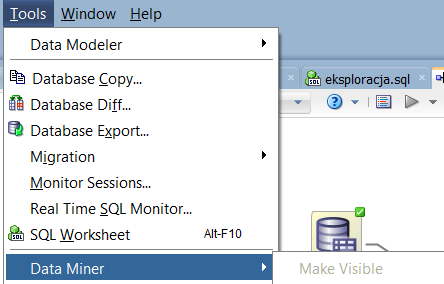
Wynik:



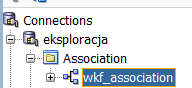
1. **Model eksploracji danych**
   1. **Model asocjacji**

Odkrywanie asocjacji jest metodą eksploracji danych pozwalającą na znajdowanie związków między występowaniem grup elementów w zbiorach danych.

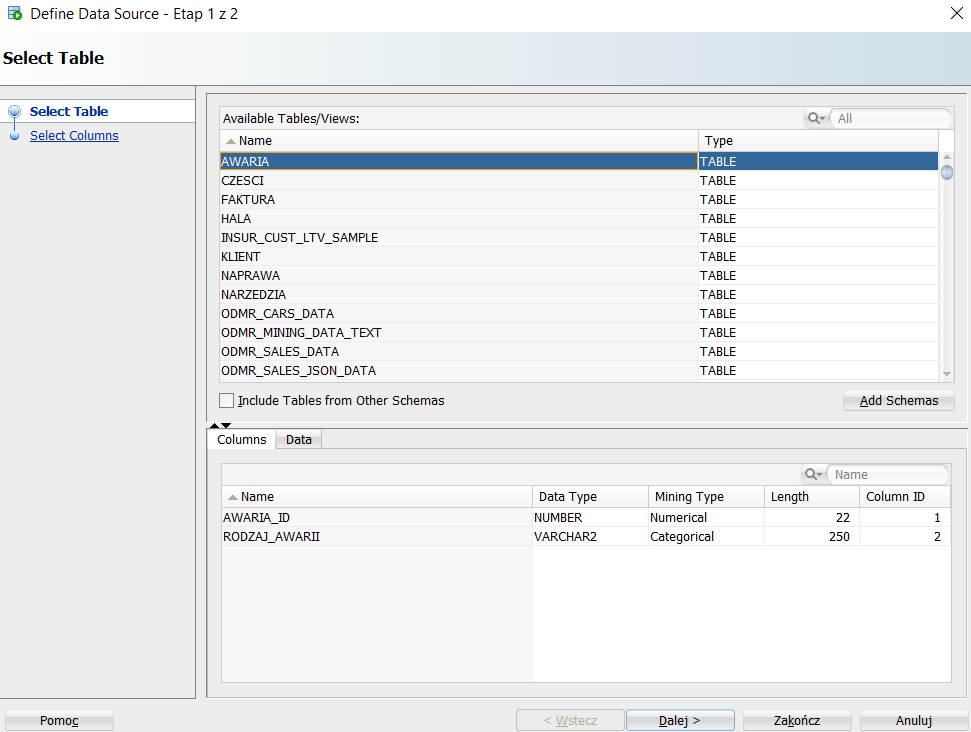
Pierwszym krokiem do zbudowania modelu odkrywania asocjacji jest uruchomienie narzędzia Data Miner w środowisku Oracle SQL developer.

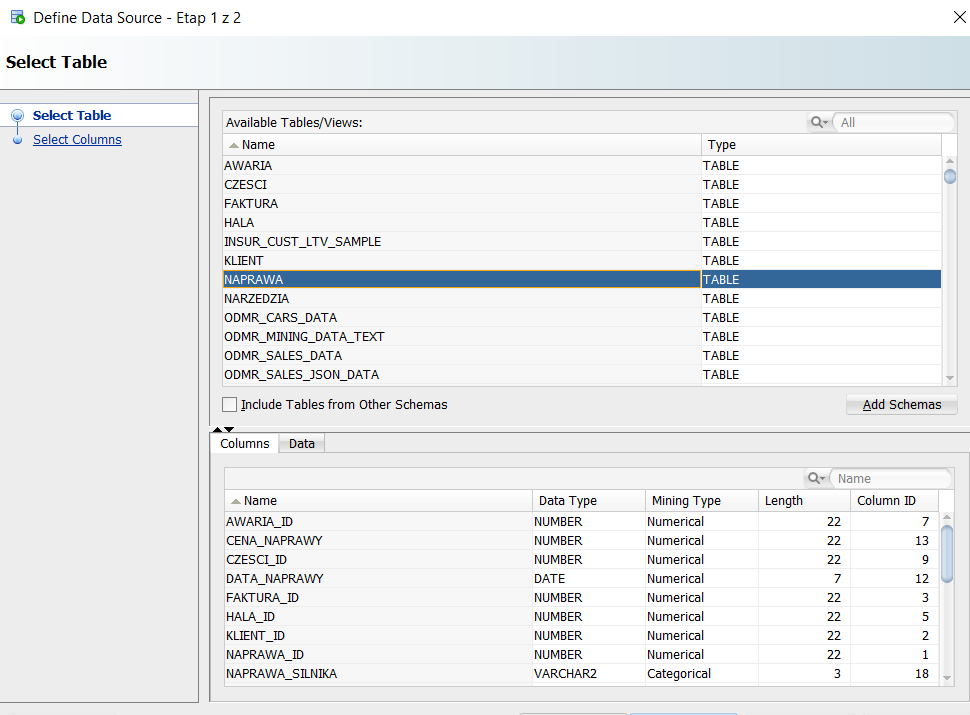
****

Następne kroki polegają na połączeniu się użytkownika, u którego znajdują się odpowiednie dane, do data minera, stworzenie nowego projektu i miejsca pracy (workspace).

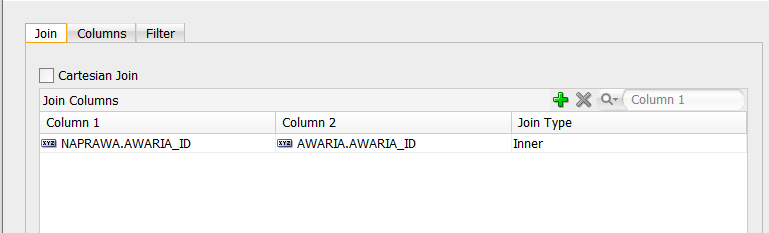


Moim celem w tym modelu jest sprawdzenie, które awarie towarzyszą wspólnie konkretnym pojazdom. W tym celu wybieram dwa obiekty Data\_source i w ich miejsce wstawiam tabele – naprawa i awaria.

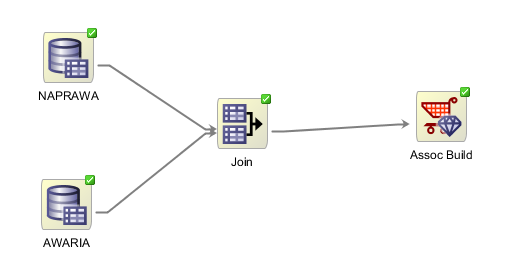




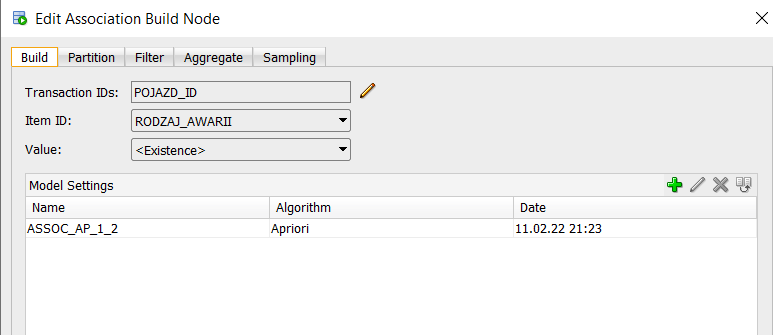
Następnie łączę obie tabele za pomocą narzędzia Join:



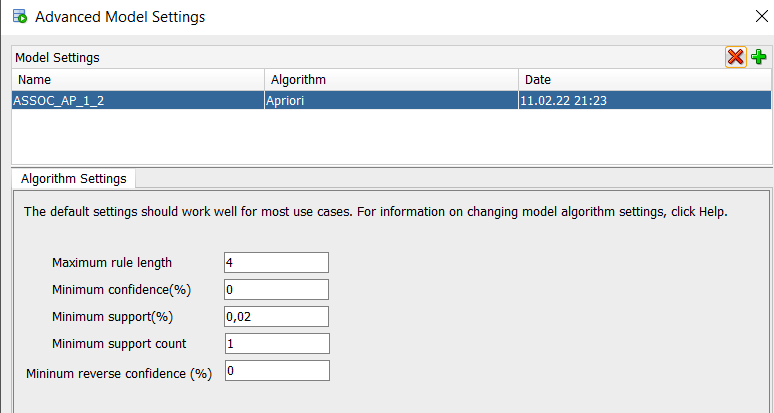
Na końcu należy jeszcze dodać narzędzie asocjacji i połączyć wszystkie obiekty strzałkami. Tak wygląda prezentowany model:



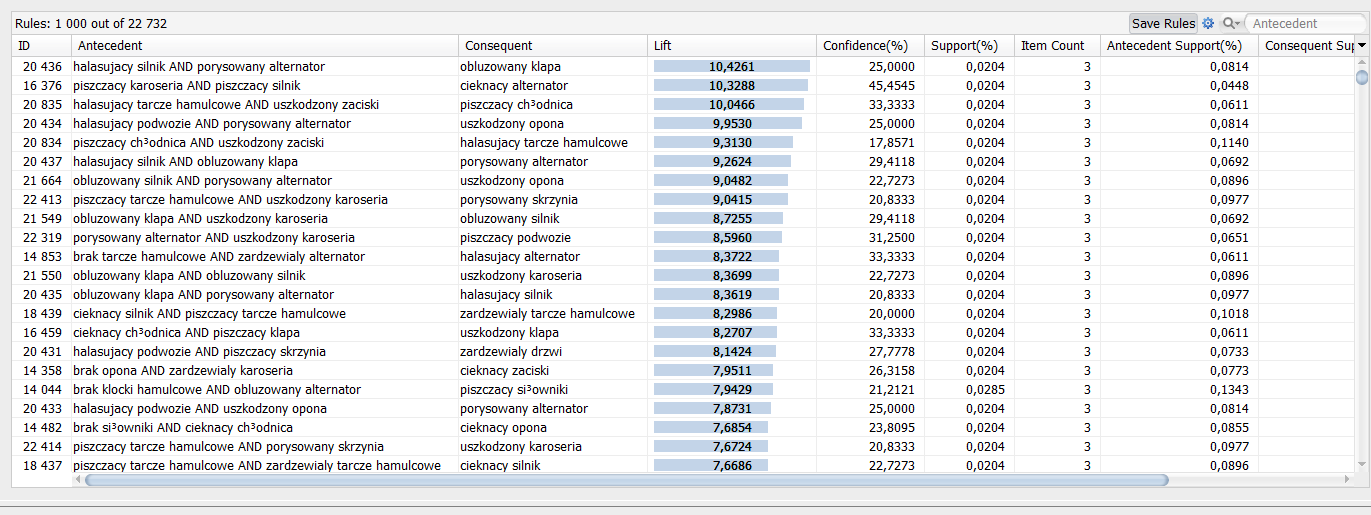
Po wybraniu asocjacji należy wybrać id transakcji – w moim przypadku będą to pojazdy, dla których będę badał awarie oraz itemID czyli wspomniana już awaria.



W opcjach zaawansowanych można wyznaczyć maksymalną długość reguł asocjacyjnych, minimalny poziom ufności i wsparcia:



Po uruchomieniu powyższego modelu można przyjrzeć się wynikom.



Po wynikach widać m.in. że hałasujący silnik i porysowany alternator często spotykane są wspólnie z obluzowaną klapą. Czy też piszcząca chłodnica i uszkodzone zaciski towarzyszą hałasującym tarczom hamulcowym.

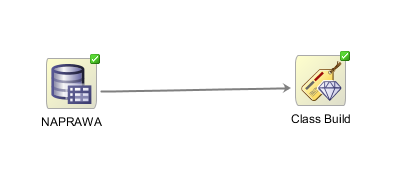
* 1. **Klasyfikacja**

Klasyfikacja jest metodą analizy danych, której celem jest predykcja wartości określonego atrybutu w oparciu o pewien zbiór danych treningowych.

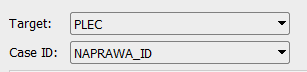
Do zbudowania modelu klasyfikacji stworzony został nowy projekt i nowy workspace



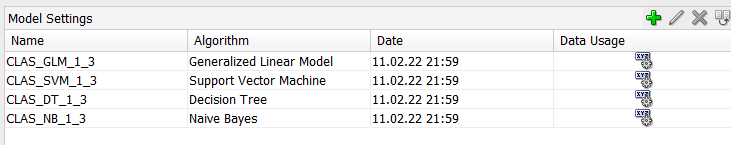
Model zbudowany jest tylko z 2 obiektów – data source, w którym znajduje się tabela i rekordy naprawy oraz Class Build pozwalającym na dokonanie klasyfikacji.



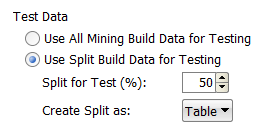
Celem klasyfikacji będzie przewidzenie płci klienta w zależności od dokonanych napraw. Co za tym idzie, zmienną celu jest płeć.



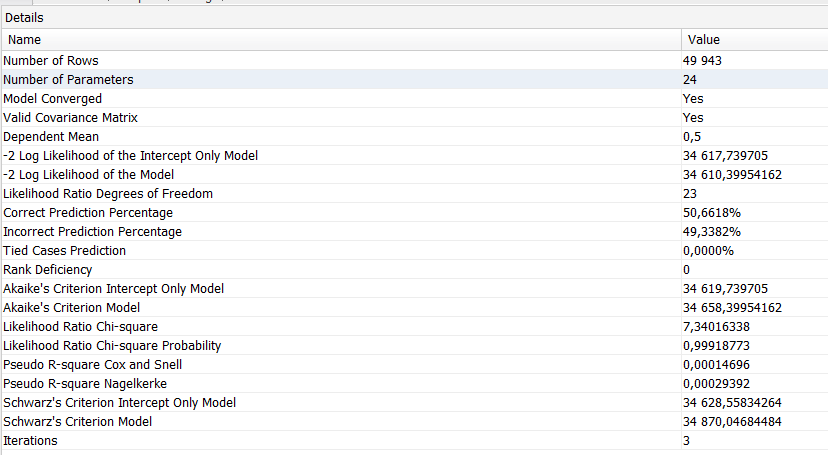
Klasyfikacja została dokonana w oparciu o 4 modele. Model naiwny Bayes’a, drzewo decyzyjne, maszynę wektorów nośnych oraz ogólny model liniowy

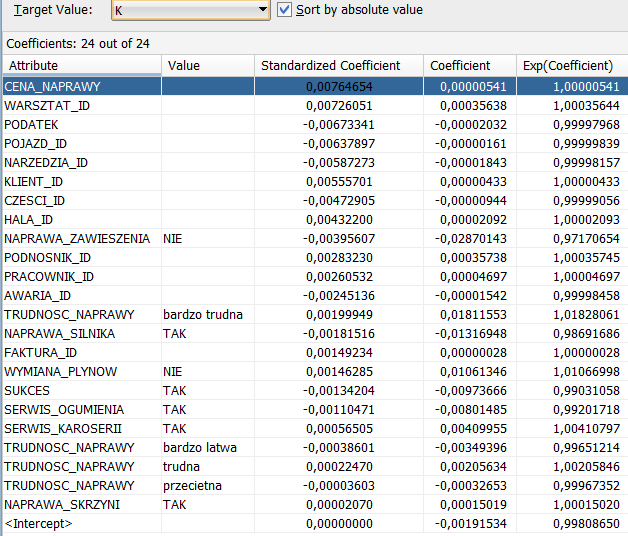


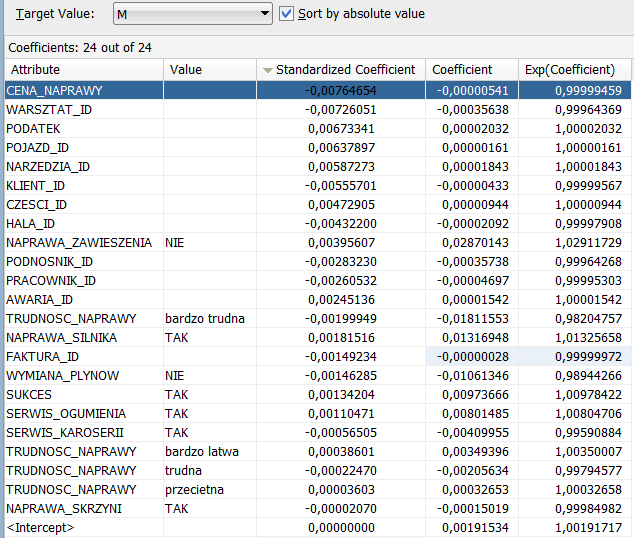
Podział na dane testowe – 50% rekordów został przeznaczony na dane testowe



Ogólny model liniowy:



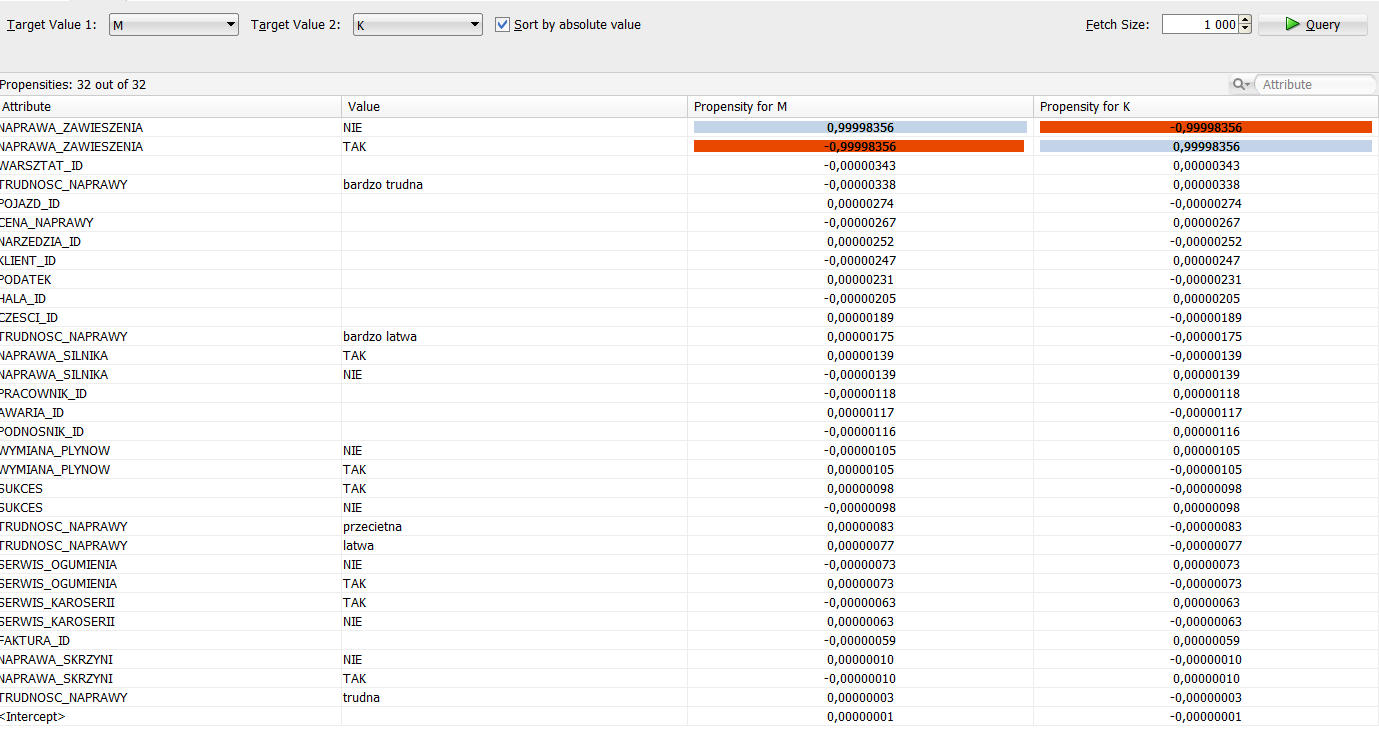




Atrybuty Standardized Coefficient i Coefficient oznaczają współczynnik korelacji i mogą przyjmować wartości w granicach <-1, 1>. Im wartość bliższa jest 0, tym mniejsza jest korelacja między klasą, a zmienną opisową. Dane zostały posortowane według standaryzowanego współczynnika korelacji malejąco

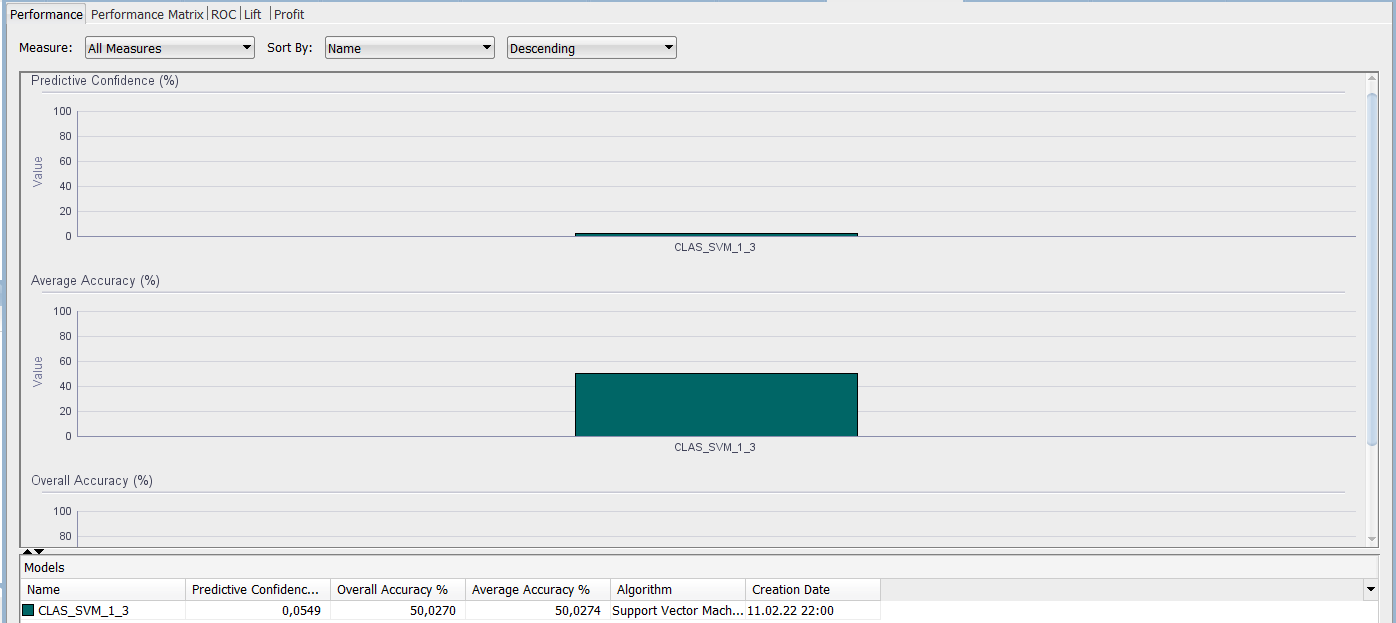
Model osiągnął precyzję na poziomie 50,6618%. Jest to słaby wynik i jest spowodowany prawdopodobnie tym, że dane w hurtowni zostały wygenerowane losowo. Oznacza to, że szansa na poprawną klasyfikację wynosi około 50%.

Maszyna wektorów nośnych



Model ten pozwala oszacować jakie zmienne miały największy wpływ na podjęcie dezycji o tym, czy klient jest kobietą czy mężczyzną. W tym przypadku Model stwierdził, że jeśli w naprawie występuje naprawa zawieszenia to klientem jest kobieta, w przeciwnym przypadku klientem jest mężczyzna. Pozostałe zmienne zostały praktycznie pominięte.

W przypadku użycia algorytmu maszyny wektorów nośnych szansa na poprawną klasyfikację również nieznacznie tylko przekroczyła 50%:

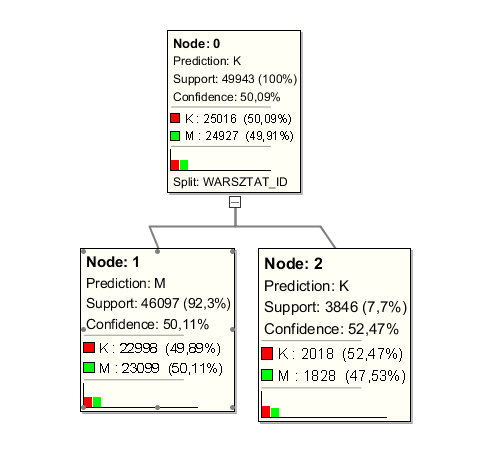


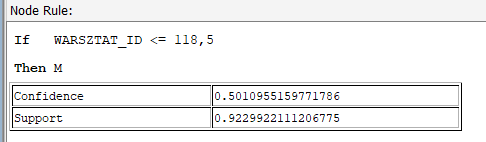
Drzewa decyzyjne

Algorytm ten buduje drzewa decyzyjne klasyfikujące dane na podstawie zbudowanych w trakcie uczenia zestawu reguł decyzyjnych. Jest to metoda przejrzysta i łatwa w odczytaniu.

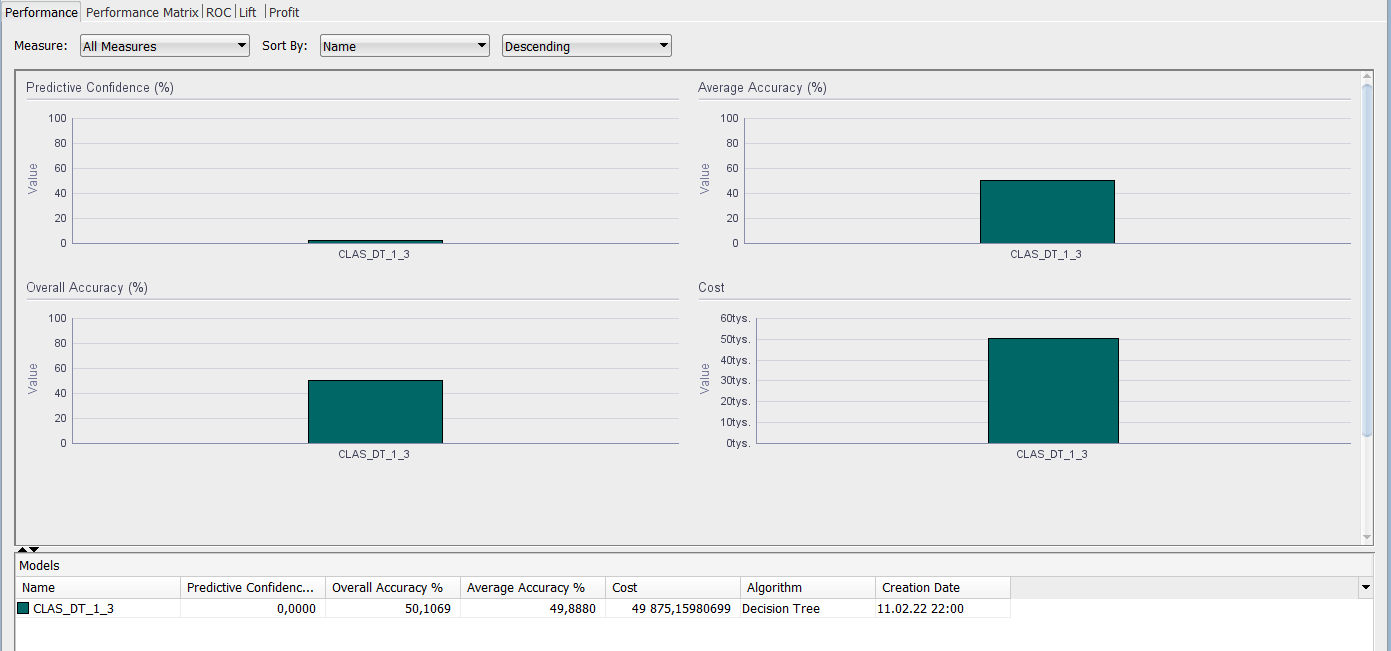
W wyniku ich użycia otrzymujemy warunki, jakie muszą spełniać zmienne opisowe aby rekord został sklasyfikowany do poprawnej klasy. W tym przypadku również losowe generowanie danych sprawiło, że wyniki nie są zbyt użyteczne. Model podzielił rekordy w zależności od zmiennej warsztat\_id. Według modelu wartość zmiennej warsztat\_id mniejsza niż 118,5 wskazuje, że naprawa została wykonana dla klienta, który jest mężczyzną.

Ufność dla obu predykcji wynosi ok. 50%.



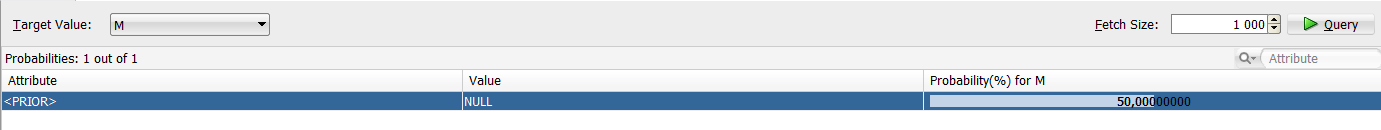


Szansa na poprawną klasyfikację ponownie nieznacznie przegracza 50%.

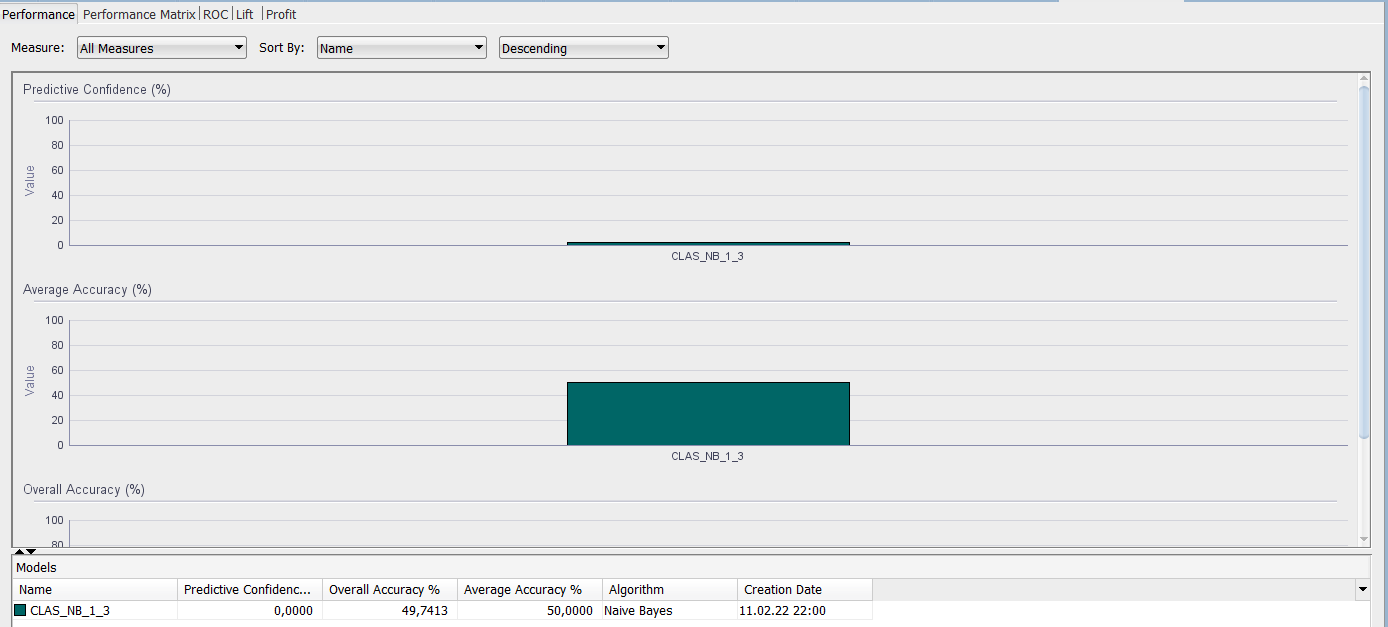


Naiwny algorytm Bayes’a

Niestety – jedyna wiadomość jaką można odczytać z tego algorytmu to prawdopodobieństwo wylosowania mężczyzny równa 50%. Ponownie może to być spowodowane losowym generowaniem danych do hurtowni.



Tym razem szansa na poprawną kwalifikację spadła poniżej 50%:



* 1. **Grupowanie**

Grupowanie jest nienadzorowaną metodą eksploracji. Oznacza to, że nie wymaga ona ustalania zmiennej celu i cała operacja wykonuje się automatycznie.

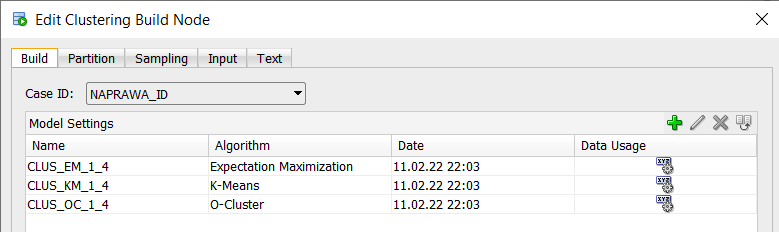
Ponownie zbudowany został nowy projekt i nowy workspace.



Model został zbudowany w oparciu o 2 obiekty – Data source posiadający rekordy z tabeli faktów – naprawa oraz obiekt grupowania.

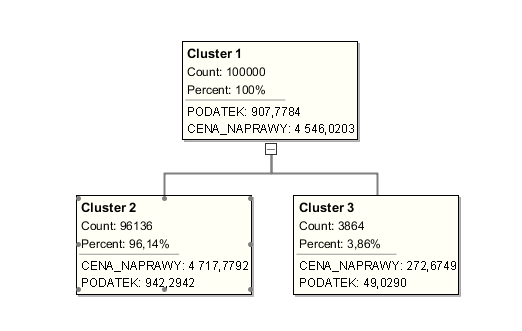


Jako Case ID ustawiony został rekord NAPRAWA\_ID będący kluczem podstawowym dla tabeli naprawa. Użyte zostały 3 algorytmy grupowania: Algorytm maksymalizacji oczekiwań, algorytm K-średnich oraz algorytm O-cluster. Rekordy zostały pogrupowane automatycznie.

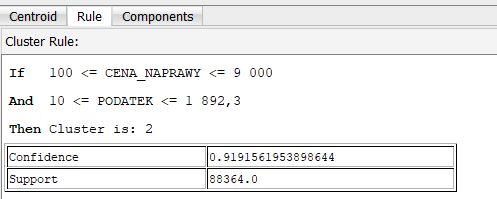


Algorytm maksymalizacji oczekiwań

Model ten buduje drzewo. W korzeniu nr 1 znajdują się wszystkie rekordy, które zostają rozbite na 2 grupy. W innych przypadkach możliwe jest dalsze rozbijanie rekordów. Jednak w przypadku mojej hurtowni rekordy zostały rozbite tylko na 2 grupy.

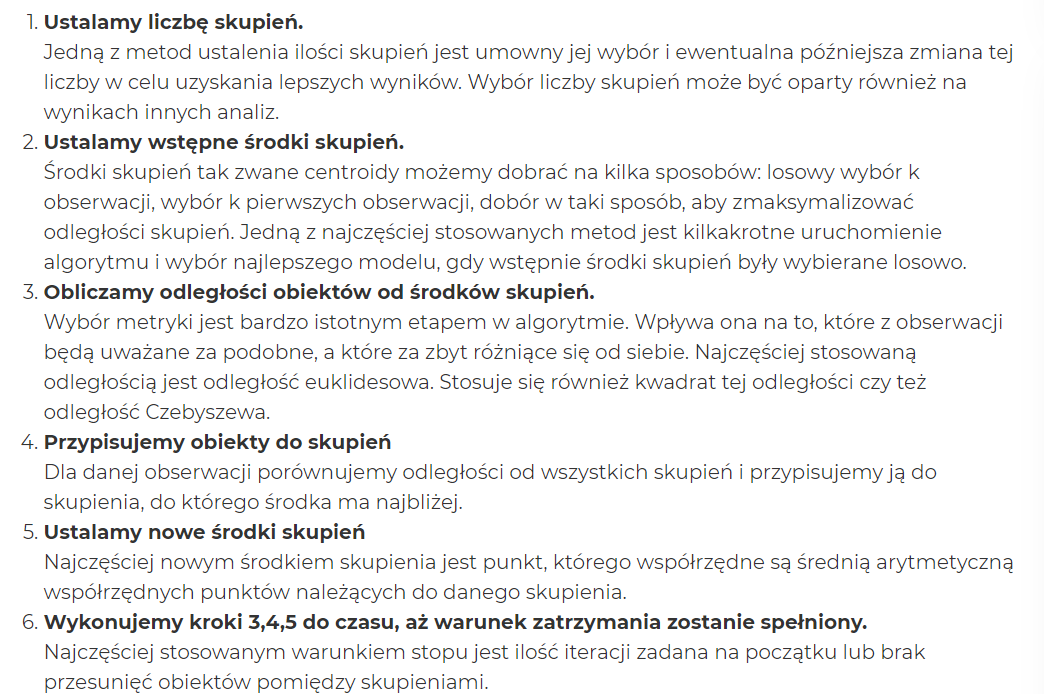


W zakładce „Rule” można sprawdzić jakie warunki musiał spełnić rekord żeby znaleźć się w konkretnej gałęzi drzewa.

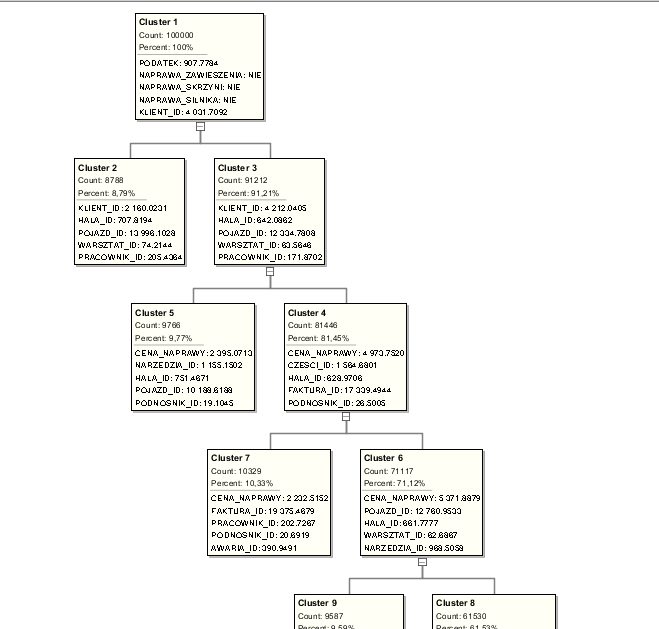


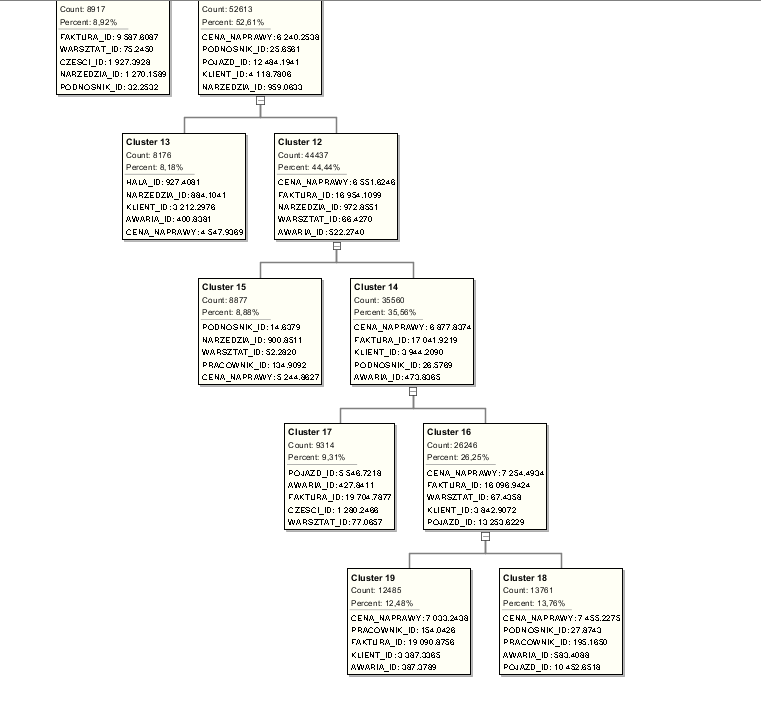
Algorytm k-średnich

Model ten również buduje drzewo, gdzie w korzeniu znajdują się wszystkie rekordy, które następnie zostają rozbite. Algorytm ten działa w oparciu o następujące kroki:

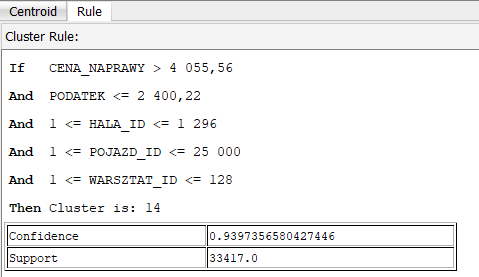


W typ przypadku algorytm podzielił rekordy na 19 gałęzi drzewa.

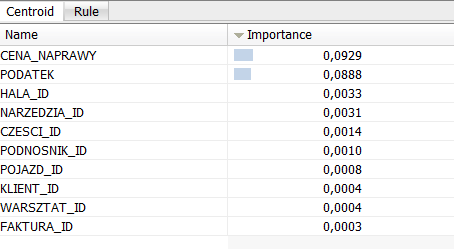




W tym przypadku również można sprawdzić jakie warunki musiał spełnić rekord by znaleźć się w odpowiedniej grupie:

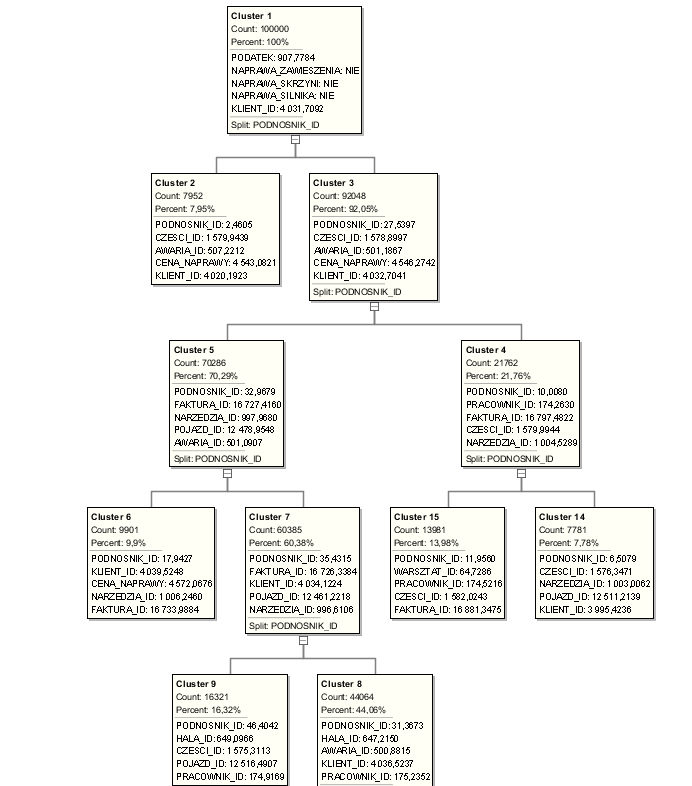


W zakładce Centroid można sprawdzić jakie atrybuty miały największy wpływ na znalezienie się rekordu w tej akurat gałęzi drzewa

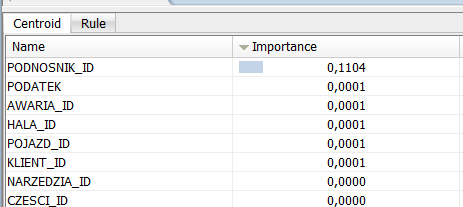


Algorytm O-cluster

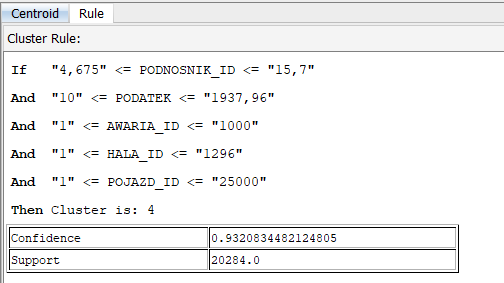
Algorytm ten, podobnie jak 2 poprzednie również buduje drzewo, gdzie w korzeniu znajdują się wszystkie rekordy i zostają w każdej następnej gałęzi dzielone na coraz mniejsze grupy



W tym przypadku również można sprawdzić w zakładce Centroid jaki atrybut miał największy wpływ na znalezienie się rekordu w konkretnej gałęzi drzewa



Można też sprawdzić w zakładce Rule jakie dokładnie warunki musi spełnić rekord by znaleźć się w konkretnej gałęzi drzewa:



1. **Wnioski**

W ramach projektu zbudowana została hurtownia danych, wygenerowane zostały dane i zaprojektowane zostały skrypty pozwalające na szybkie załadowanie danych do hurtowni. Napisanych zostało 18 zapytań, po 3 z kategorii Cube, Rollup, Grouping Sets, Partycje obliczeniowe, Rankingi, Okna ruchome. Zapytania pozwalają na wyciągnięcie ciekawych wniosków z hurtowni takich jak: wyznaczenie najlepiej zarabiających warsztatów, sprawdzenie w jaki dzień miesiąca i w jaki miesiąc zyski są największe czy też wyznaczenie pracowników którzy w konkretnym roku zarobili najwięcej.

Zostały zbudowane również 3 modele eksploracji danych. Model asocjacji pozwolił przewidzieć jaki rodzaj awarii jest powiązany z innym dla konkretnego samochodu. Reguły asocjacyjne wyróżnione w tym modelu miały jednak niskie wsparcie spowodowane najprawdopodobniej losowym generowaniem danych.

Model klasyfikacji miał za zadanie przewidzieć płeć klienta na podstawie zmiennych. Wynik nie był jednak zbyt wiarygodny. Szansa na poprawną klasyfikację w 3 z 4 użytych algorytmów nieznacznie przekroczyła 50% co znaczy, że przewidywać można by równie dobrze rzucając monetą. W tym przypadku również najprawdopodobniejszym powodem takiej niewielkiej szansy na poprawną klasyfikację są losowo generowane dane.

Ostatnią metodą eksploracji danych było grupowanie. Za pomocą 3 algorytmów rekordy z tabeli faktów hurtowni zostały pogrupowane na różne grupy. Można było sprawdzić jaka miara tabeli miała największy wpływ na grupowanie w danej gałęzi drzewa oraz jakie warunki rekord musiał spełnić aby znaleźć się w konkretnej gałęzi drzewa.