Dane do analizy

- Katalog *dane: DRUG1n, customer_dbsave.sav, tree_credit.sav , pm_customer_train1.sav , pm_customer_train2.sav, pm_customer_train3.sav, GOODS1n, GOODS2n, Baskets1n,*

Ćwiczenie 1
Przewidywanie efektywności leczenia

Wczytywanie danych tekstowych

W zakładce „Źrodła" wybieramy „Plik separowany" dwukrotnie klikając i wybieramy *DRUG1n*.
W zakładce „Typy" „Odczytaj wartości"

Pola w pliku źródłowym:

Age	Wiek (ilościowa)
Sex	Płeć (M lub F)
BP	Ciśnienie krwi: HIGH, NORMAL lub LOW
Cholesterol	Poziom cholesterolu: NORMAL lub HIGH
Na	Poziom sodu (ilościowa)
K	Poziom potasu (ilościowa)
Drug	Lek, który podziałał na pacjenta (zmienna docelowa)

Oglądanie danych - dodanie tabeli

Zobacz dane wejściowe: włącz do strumienia węzeł „Tabela" z zakładki „Wyniki i „Wykonaj".

Tworzenie grafu rozkładu

W celu znalezienia odpowiedź na pytanie*, jaka część pacjentów odpowiedziała pozytywnie na terapię danym lekiem*, można wykorzystać węzeł „Rozkładu" z zakładki „Wykresy" i łączymy ze źródłem.
Klikamy dwukrotnie w celu ustalenia opcji. W polu „Zmienna" wybieramy „Drug" i uruchamiamy wykonanie strumienia.
Analiza grafu pozwala zaobserwować „kształt" danych: okazuje się, że pacjenci odpowiadali pozytywnie najczęściej na lek Y a najrzadziej na B i C.
Można również uruchomić węzeł „Audyt danych" z zakładki „Wyniki".

Tworzenie wykresu rozrzutu

Chcemy zidentyfikować czynniki wpływają na „Drug" (zmienną docelową). Wiadomo, że koncentracja sodu i potasu we krwi są istotnymi czynnikami. Ponieważ obydwa przyjmują wartości liczbowe, możemy utworzyć wykres rozrzutu sodu wobec potasu używając kategorii leku jako nałożenia.

Wstaw wykres „Rozrzutu” z zakładki „Wykresy” i połącz ze źródłem. Kliknij dwukrotnie w celu ustalenia parametrów. Jako „Zmienna X” wybierz „Na”, a „Zmienna Y”: K , a „Nałożenie”: „Drug”. Uruchom strumień.
Wykres pokazuje progi powyżej których poprawnym lekiem jest Y i poniżej którego poprawnym lekiem nie jest nigdy Y. Próg ten, to stosunek sodu do potasu.

Tworzenie wykresu sieciowego

Ponieważ kilka pól przyjmuje wartości ze zbioru kategorii, można spróbować narysować wykres sieciowy, który obrazuje zależności między różnymi kategoriami. Wstaw węzeł „Sieciowy” z zakładki „Wykresy” i połącz ze źródłem. Wybierz zmienne: „BP” oraz „Drug” i „Wykonaj”. Z wykresu widać, że „drugY” jest związany ze wszystkim trzema poziomami ciśnienia krwi. Aby zanalizować inne lekarstwo, można ukryć „drugY”: prawy przycisk i wybór: „Ukryj i zaplanuj ponownie” („drugY” i wszystkie jego powiązania zostaną ukryte). Teraz widać, że tylko leki A i B są związane z wysokim ciśnieniem krwi. Tylko C i X są związane z niskim. A normalne ciśnienie jest związane tylko z lekiem X. W dalszym ciągu nie wiadomo na jakiej podstawie dokonać wyboru między lekiem A i B lub między C i X dla danego pacjenta. W takim przypadku pomoże modelowanie.

Wyliczanie nowego pola

Ponieważ stosunek sodu do potasu wydaje się być czynnikiem wpływającym na zastosowanie leku Y, można wstawić nowe pole zawierające taką wartość dla każdego rekordu. Pole takie może okazać się przydatne później do zbudowania modelu przewidującego, kiedy użyć każdy z pięciu lekarstw.
Wstaw węzeł „Wyliczanie” z zakładki „Zmienne”. Kliknij dwukrotnie, wpisz jako „Zmienna wyliczana”: „Na_to_K”, jako formułę wpisz „Na/K”.
Możemy teraz sprawdzić rozkład nowej zmiennej wstawiając węzeł „Histogram” z zakładki „Wykresy” i łącząc go z węzłem wyliczanym „Na_to_K”. Kliknij dwukrotnie i wpisz „Na_to_K” jako zmiennę do rysowania i „Drug” jako „Nałożenie”. Analizując otrzymany graf można dojść do wniosku, że jeżeli wartość „Na_to_K” jest powyżej 15, to wybierany jest lek Y.

Budowanie modelu

Analizując dane można postawić pewne hipotezy (stosunek sodu do potasu wydaje się mieć wpływ na wybór terapii podobnie na ciśnienie krwi). Nie możemy jednak w pełni określić wszystkich zależności. Spróbujemy zbudować model klasyfikacyjny C5.0.
Ponieważ używamy pola wyliczanego „Na_to_K”, możemy odfiltrować oryginalne pola „Na” i „K” tak by nie były wykorzystywane w modelu dwukrotnie.
Można to zrobić używając węzła „Filtrowanie” z zakładki „Zmienne” podłączając go do węzła „Na_to_K”. Klikamy dwukrotnie na węzeł „Filtrowanie” i w zakładce „Filtr” klikamy strzałki w polach „Na” i „K” – pojawi się przekreślenie oznaczające, że te pola zostaną wyłączone.
Wstaw następnie węzeł „Typy” z zakładki „Zmienne” i połącz go z węzłem „Filtrowanie”. Kliknij dwukrotnie i w zakładce „Typy” ustaw „Kierunek” dla zmiennej „Drug” na „Przewidywana” wskazując w ten sposób pole „Drug” jako pole, które chcemy przewidywać. Pozostałe zmienne pozostaw – będą one predykatorami.
Aby zbudować model wstaw węzeł „C5.0” z zakładki „Modelowanie” i połącz z węzłem „Typy” i uruchom.

Wygenerowany model pojawi się w prawym górnym rogu w zakładce „Modele". Aby go przeglądnąć należy kliknąć prawym przyciskiem i wybrać „Przeglądaj".

Użycie węzła analizy

Teraz można oszacować dokładność otrzymanego modelu. Po pierwsze należy wstawić wygenerowany model do strumienia (prawy przycisk na wygenerowanym modelu i wybrać „Dodaj do strumienia" wcześniej wybierając węzeł „Typy". Następnie dodać węzeł „Analiza" z zakładki „Wyniki" (wcześniej wybierając węzeł „Drug") i uruchomić węzeł. Pojawiają się wyniki, które mówią, że na takim zbiorze danych model przewiduje poprawnie wybór leku dla prawie każdego rekordu. Oczywiście na rzeczywistego zbioru ta dokładność będzie dużo niższa.

Ćwiczenie 2. Badanie predyktorów

Wybór cech (*feature selection*) pozwala zidentyfikować atrybuty, które mają największy wpływ na predykcję atrybutów wyjściowych.

Dane *customer_dbsave.sav* zawierają informacje o odpowiedziach na pytania ankietowe udzielonych przez 5000 klientów firmy telefonicznej. Dane zawierają informacje o wieku klienta, zatrudnieniu, dochodach i statystyki wykorzystania telefonu. Trzy „docelowe" pola wskazują, czy klient odpowiedział na zapytania ofertowe. Firma chce wykorzystać te dane do wspomagania predykcji którzy klienci odpowiedzą w przyszłości na podobne oferty.

Przykład wykorzystuje tylko jedną ofertę jako cel. Wykorzystany zostanie model drzewa decyzyjnego CHAID: bez wyboru cech i z wyborem cech (10 najlepszych predykatorów). W drugim przypadku otrzymane wyniki okażą się efektywniejsze.

Budowanie strumienia

W pustym oknie wstaw źródło „Plik Statistica" i jako źródło wskaż plik customer_dbsave.sav.
Wstaw węzeł „Typy" z zakładki „Zmienne" i zmień kierunek dla „response_01" na „Przewidywana". Dla „custid", „response_02", „response_03" ustaw kierunek na „Brak" .
Wstaw węzeł „Wybór predyktorów" z zakładki „Modelowanie" wskazując wcześniej węzeł „Typy" jako źródło. Uruchom strumień.
Kliknij otrzymany model (w prawym górnym rogu) i prawym przyciskiem myszy wybierz „Przeglądaj". Górny panel pokazuje pola, które zostały uznane za użyteczne w predykcji.

Wstaw zbudowany model do strumienia wskazując wcześniej węzeł „Typy" i kliknij na niego dwukrotnie. Wybierz 10 pierwszych predyktorów.
Aby porównać wyniki bez wyboru predykatorów trzeba wygenerować dwa modele: jeden który użyje wybór i drugi, który tego wyboru cech nie wykorzysta. Wstaw dwa węzły „CHAID'" z zakładki „Modelowanie": jeden połącz do węzła „Typy" (zmień jego nazwę na „All input fields", drugi do wygenerowanego „response_01" (zmień jego nazwę na „Top 10 fields") . Uruchom węzeł „All input fields" – zauważ jak długo się wykonuje. Wykonaj to samo dla drugiego węzła.
Drugi model wykonał się szybciej (różnice te na pewno będą bardziej zauważalne dla większej liczby danych). Drugie drzewo zawiera również mniej węzłów niż pierwsze. Jest łatwiejsze do interpretacji i zrozumienia.

Ćwiczenie 3. Tworzenie profili klientów

Przykład wykorzystuje dane *Baskets1n.*

1. Wstaw „Plik separowany" i zdefiniuj jako źródło plik „Baskets1n".
2. Wstaw węzeł „Typy".
3. Następnie wstaw węzeł „Tabela".
4. W węźle „Typy" ustal typ zmiennej „Cardin" na „Brak" (ponieważ każdy numer karty lojalnościowej występuje tylko raz i nie będzie wykorzystywany w modelowaniu).
5. Ustaw „Jakościowa" jako typ dla atrybutu „Sex".
6. Uruchom strumień – zbiór zawiera 18 pól, każdy rekord reprezentuje koszyk:
7. Podsumowanie koszyka:
    - cardid – nr karty lojalnościowej
    - value – wartość koszyka
    - pmethod – sposób płatności

Charakterystyka klienta:

- sex, homeown (czy klient posiada dom), income, age

Zawartość koszyka – flaga określająca czy produkt z danej kategorii znalazł się w koszyku

- fruitveg, freshmeat, dairy, cannedveg, cannedmeat, frozenmeal, beer, wine, softdrink, fish, confectionery

8. W węźle "Typy" ustaw "Kierunek" dla wszystkich produktów na „Obydwie" (atrybut może być zarówno wejściem jak i wyjściem w modelu), a wszystkie pozostałe na „Brak".
9. Do węzła „Typy" podłącz węzeł „Apriori" z zakładki „Modelowanie", edytuj go i zaznacz „Tylko wartości prawda dla flag" i wykonaj węzeł Apriori.
10. W prawym górnym rogu dostajemy model, który możemy przeglądać.
11. Do węzła „Typy" podłącz węzeł „Sieciowy" z zakładki „Wykresy", wybierz wszystkie pola z produktami, ustaw „Pokaż tylko flagi prawdy" i uruchom węzeł.
12. Naciśnij dwie żółte strzałki i wybierz dla „Widok sieci": „Styl wyróżnia silne/normalne/słabe łącza". Ustaw, aby silne zależności były powyżej 100 a słabe poniżej 90.
13. W wyniku widać 3 grupy klientów:
    - tych którzy kupują ryby i owoce i warzywa („zdrowo odżywiający się")
    - tych którzy kupują wino i słodycze
    - tych którzy kupują piwo, mrożonki i warzywa w puszce („piwo, fasola i pizza")

Wyszukiwanie profili grup klientów

1. Do 3 znalezionych grup klientów chcemy się dowiedzieć teraz coś więcej na temat charakterystyki tych klientów (profil demograficzny)
    - można to uzyskać zaznaczając każdego klienta flagą przynależności do poszczególnych grup i wykorzystać metodę indukcji reguł (C5.0) dla każdej z flag.
2. Najpierw należy ustawić flagę dla każdej z grup -można ją wygenerować automatycznie przy wykorzystaniu wygenerowanego wykresu sieciowego.
3. Kliknij prawym przyciskiem na link między „fruitveg" i „fish" i wybierz „Utwórz węzeł wyliczeń dla łącza". Kliknij dwukrotnie na powstały węzeł (zmienną), żeby zmienić nazwę na „healthy".
4. Powtórz procedurę dla połączenia „wine" z „confectionery", a powstałe pole nazwij „wine_chocs".

5. Dla trzeciej grupy upewnij się, że żadna krawędź nie została wybrana, wybierz następnie 3 linki w trójkącie „cannedveg", „beer" i „frozenmeal" trzymając klawisz „Shift" przy wyborze każdej z 3 krawędzi. Następnie wybierz z menu „Generuj": „Węzeł wyliczeń „And"". Powstały węzeł nazwij „beer_beans_pizza".
6. W celu wyliczenia profili użytkowników, połącz węzeł „Typy" z tymi trzema otrzymanymi węzłami wyliczanymi w serii, a następnie wstaw nowy węzeł „Typy".
7. W nowo wstawionym węźle ustaw „kierunek" wszystkich zmiennych na „brak" z wyjątkiem „value", „pmethod", „sex", „homeown", „income" i „age" – ich kierunek powiniem być ustawiony na „Wejście" i wybrana grupa (np. „beer_beans_pizza") na „Przewidywana".
8. Wstaw węzeł modelujący „C5.0", w którym ustaw „Typ wyjściowy" na „Zestaw reguł" i uruchom.
9. W modelu dostajemy regułę w rodzaju:

   Reguła 1 dla P

   jeżeli income <= 16900

   I sex = M

   to P
10. Zastosuj podobne procedury do dwóch pozostałych grup poprzez wybór ich jako wyjście.

Ćwiczenie 4. Klasyfikacja - scoring kredytowy w oparciu o algorytm CHAID

Przykład wykorzystuje dane *tree_credit.sav:*

Credit_rating	Credit rating: 0=Bad, 1=Good, 9=missing values (zm. docelowa)
Age	Age in years
Income	Income level: 1=Low, 2=Medium, 3=High
Credit_cards	Number of credit cards held: 1=Less than five, 2=Five or more
Education	Level of education: 1=High school, 2=College
Car_loans	Number of car loans taken out: 1=None or one, 2=More than two

1. Wstaw odpowiedni węzeł do strumienia
2. Ustaw atrybut *credit_rating* jako zmienną predykcyjną
3. Wstaw węzeł modelujący CHAID
   - w zakładce "Zmienne" powinno być wybrane ustawienie "Użyj wstępnie zdefiniowanych ról"
   - w zakładce "Budowanie":
     - "Zbudować pojedyncze drzewo" ->"Utwórz gotowy model"
   - w zakładce "Kryteria zatrzymania":
     - "Wartość bezwzględna" ("Minimum rekordów w gałęzi nadrzędnej"=400, "Minimum rekordów w gałęzi podrzędnej"=200)
   - uruchom węzeł
4. Oglądnij model:
   - w postaci regułowej: które predyktory są najsilniejsze? jakie są różnice m. nimi? które nie mają znaczenia?
   - w postaci drzewa:
     - jak wygląda węzeł 2 (low income) i czy wszystko jest w porządku?
     - podobnie węzeł 1 (high income)? jak zmienia się dokładność po dalszym podziale tego węzła?
     - co z dokładnością w węźle 3 (medium income)? co zmienia dalszy podział tego węzła?
5. Ocena dokładności modelu:

- włącz węzeł "Tabela" do strumienia (za modelem) i uruchom
- w "$R-Credit_rating" - wartość predykowana
- w "$RC-Credit_rating" - wartość confidence (dokładność predykcji od 0 do 1)
- wartość szacowania: węzeł "Analiza" (dołącz go do strumienia i uruchom)
- przetestuj model na innym pliku bez zmiennej klasowej (stwórz go)

Ćwiczenie 5. Przewidywanie efektów sprzedaży (C&RT)

Przykład wykorzystuje dane GOODS1n (plik separowany) (GOODS2)
Class kategoria produktu
Cost cena jednostkowa
Promotion wskaźnik ilości czasu poświęconego konkretnej promocji
Before wpływy przed promocją
After wpływy po promocji

1. Wstaw źródło
2. Wprowadź nowe pole liczące % wzrostu przychodu i nazwij je "Increase"
3. Sprawdź jak wygląda rozkład zmiennej "Increase" a jak wykres wzrostu przychodów względem kosztów promocji ("class" jako nałożenie)
4. Ustaw "Increase" jako "przewidywalną" , a "After" jako "brak"
5. Podepnij węzeł C&RT
   - ustaw maksymalną głębokość = 7
   - nie przycinaj
6. Podepnij węzeł "Analiza" i oglądnij wyniki
7. Zrób wykres "Increase" względem "$R-Increase"
8. Sprawdź dokładność modelu na zbiorze GOODS2n
9. Jak można dostroić model?

Ćwiczenie 6. Przewidywanie efektów sprzedaży (Auto klasyfikacja)

Przykład wykorzystuje dane *pm_customer_train1.sav* (historyczne dane o ofertach składanych klientom)*:*
Campaign cztery kampanie promocyjne kont klienckich (np. 2=Premium account)
Response czy oferta zostala zaakceptowana (0=nie, 1=tak) - zmienna docelowa

1. Wczytaj źródło
2. Ustaw *response* jako przewidywana, a jej "Poziom pomiaru" na flaga
3. Dla atrybutów: customer_id, campaign, response_date, purchase, purchase_date, product_id, Rowid, X_random ustaw "Rola" na brak
4. Odczytaj wartości
5. Zastąp identyfikatory kampanii numerami: dla atrybutu campaign w polu "Wartości" wybierz "Określ"
   - w "Poziom pomiaru" wybierz "Nominalna i wpisz:
     1 Standard account
     2 Premium account
     3 Gold account
     4 Platinum account
6. Podłącz węzeł "Tabela" i zobaczy wyniki (wybierz "Wyświetl etykiety zmiennej i wartości")
7. Analiza każdej kampanii po kolei: wstaw węzeł "Selekcja" i wybierz najliczniejszą grupę "Premium" (2)
8. Generowanie i porównanie modeli:

- wstaw węzeł "Autoklasyfikacja" i wybierz "Całkowitą dokładność" jako pomiar modeli ("Ranguj model według")
- "Liczbę modeli do wykorzystania" ustaw na 3
- Zakładka "Zaawansowane" - 11 modeli do wyboru (odznacz najdłużej uczące się: "Analizę dyskryminacyjną" i "SVM")
- W zakładce "Ustawienia" w "Metoda zespolenia" wybierz "Głosowanie ważone ufnością"
- Uruchom węzeł
- Przeanalizuj modele
    - czy jest któryś znacząca lepszy ("ogólna dokładność")?
    - kliknij na dowolny model w celu uzyskania szczegółów
    - podłącz węzeł "Analiza" i uruchom - pojawia się błąd zastosowania zespołowego modelu (*$XF-response*) - jak się ma do dokładności pojedynczych modeli?

9. Przeanalizuj pozostałe kampanie

Ćwiczenie 7. Przewidywanie efektów sprzedaży (SLRM - Self-Learning Response Model)

Przykład wykorzystuje dane *pm_customer_train1.sav (pm_customer_train2.sav, pm_customer_train3.sav)*

1. Wstaw źródło do strumienia
2. Wstaw węzeł "Wypełnianie"
    - wybierz zmienną do konwersji ("campaign"), "zamień="zawsze" oraz "zamień na" ="to_string(campaign)
3. W węźle "Typy":
    - wyeliminuj "customer_id", "response_date", "purchase_date", "product_id", "Rowid", "X_random"
    - jako przewidywalne ustaw "campaign" i "response" (dla niej ustaw "Flaga")
4. Przekoduj numery kampanii (1-4) na "Mortgage", "Car loan", "Savings", "Pension"
    - dodaj węzeł "Rekodowanie":
        - "Rekoduj na"="Istniejąca zmienna"
        - "Rekoduj zmienną"="campaign"
        - wybierz "Uzyskaj" i wpisz odpowiednie kodowanie
5. Dodaj węzeł "SLRM"
    - "Zmienna przewidywana"="campaign", "Przewidywana zmienna odpowiedzi"="response"
    - w zakładce "Ustawienia": "Maksymalna liczba predykcji na rekord"=2
    - wybierz „Weź pod uwagę rzetelność modelu”
    - Uruchom węzeł
6. Otwórz model
    - Zakładka "Model"
        - szacowana dokładność predykcji dla każdej oferty i relatywna ważność każdego predyktora w szacowaniu modelu
        - korelacja każdego predyktora ze zmienną docelową: "Widok"="powiązanie z odpowiedzią"
        - konkretna oferta dla której była robiona predykcja: "Widok" (po lewej stronie)
7. Zamknij okienko

8. Podłącz plik dane *pm_customer_train2.sav* do węzła "Wypełnianie"
9. W węźle SLRM w zakładce "Model" wybierz "Kontynuuj uczenie istniejącego modelu"
10. Uruchom węzeł
11. Aby zobaczyć szczegóły kliknij 2-krotnie wynik - widać jak zmieniała się dokładność predykcji dla każdego modelu
12. Znów podłącz nowy plik (dane *pm_customer_train3.sav*)
13. Powtórz czynności (co się stało ze średnia dokładnością?)
14. Dodaj węzeł "Tabela" do trzeciego wygenerowanego modelu i go wykonaj
15. Sprawdź dokładności klasyfikacji:
    - zaobserwuj wartości $SC-campaign-1, $SC-campaign-2, $SC-campaign-3

Ćwiczenie 8. Badanie dokładności klasyfikacji

1. Wstaw źródło (plik separowany) i wczytaj dane "adult_data"
2. Wstaw węzeł "Typy" o ustaw "workclass" ja "przewidywana"
3. Wstaw węzeł "Podział" i ustaw 80:20
4. Wstaw 2 węzły "Selekcja"
    a. w górnym ustaw warunek selekcji Podział="1_Uczenie"
    b. w dolnym: "2_Testowanie"
5. Do górnego węzła podepnij węzeł "C5.0" a następnie "Analiza" - uruchom o sprawdź dokładność
6. Do dolnego podepnij zbudowany model, a następnie węzeł "Analiza" - uruchom i porównaj wyniki