Zrozumienie i przygotowanie danych – przykładowe techniki umożliwiające zrozumienie i przygotowanie danych, kolejne przykłady[1]

I. Odkrywanie zależności pomiędzy zmiennymi

Odkrywanie korelacji pomiędzy zmiennymi w zbiorze danych jest jedną z najciekawszych i najważniejszych czynności wykonywanych w ramach EDA. Jest także ważne z punktu widzenia przygotowania danych do dalszej analizy, gdyż są przypadki, kiedy duża liczba skorelowanych atrybutów może zaburzyć wyniki.

Popularne narzędzia wykorzystywane do wykrywania skorelowanych zmiennych to macierze wykresów punktowych i regresja.

1. Analiza zmiennych skorelowanych z wykorzystaniem macierzy wykresów punktowych i regresji

1. Po uruchomieniu systemu i wybraniu odpowiedniej przestrzeni roboczej wybierz z repozytorium węzłów węzeł *IO→Read→File Reader* i przenieś go na przestrzeń roboczą projektu.

2. Pod prawym przyciskiem myszy po kliknięciu na wybrany węzeł znajduje się menu kontekstowe. Wybierz *Configure...*

3. W polu *valid URL* wybierz plik *churn.txt*.

4. Przenieś węzeł *Views→Local (Swing)→Scatter Matrix (local)* na przestrzeń roboczą projektu i połącz wyjśce węzła *File Reader* z wejściem węzła *Scatter Matrix*. Zapoznaj się z opisem węzła *Scatter Matrix*.

5. Wybierając z menu kontekstowego węzła *Scatter Matrix* opcję *Execute and Open Views* obejrzyj korelacje pomiędzy minutami rozmów, liczbą rozmów i opłatą za rozmowy w ciągu dnia (*day mins, day calls, day charge*).

6. Przenieś węzeł *Analytics→Mining→Linear/Polynomial Regression→Linear Regression (Learner)* na przestrzeń roboczą projektu i połącz wyjście węzła *File Reader* z wejściem węzła *Linear Regression (Learner)*. Zapoznaj się z opisem węzła *Linear Regression (Learner)*.

[1] Przykłady zostały opracowane na podstawie: Kursu IBM: Introduction to IBM SPSS Modele rand Data Mining (Student Guide) oraz książki Daniela T. Larose „Odkrywanie wiedzy z danych" Wprowadzenie do eksploracji danych. Wydawnictwo Naukowe PWN, Warszawa 2006.

7. Skonfiguruj węzeł *Linear Regression Learner*. Jako atrybut celu podaj *Day charge* (całkowita opłata za rozmowy w ciągu dnia), jako atrybut wejściowy pozostaw jedynie *Day mins*.

8. Uruchom węzeł za pomocą opcji *Execute and Open Views*. Pamiętaj, że regresja wyznacza współczynniki a_i równania $y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots$, gdzie y to atrybut celu, a x_i to wartości atrybutów wejściowych. W tabelce będącej podstawowym widokiem dla węzła *Linear Regression Learner* wyznaczone współczynniki a_i znajdują się w kolumnie „Coeff.", kolumna „Variable" oznacza atrybut, przy którym znajdzie się współczynnik (pozostałe kolumny określają siłę korelacji między danym atrybutem a atrybutem celu). Atrybut *Intercept* jest specjalny i oznacza wiersz dla współczynnika a_0.

9. Na podstawie informacji z poprzedniego punktu zastanów się, jaka jest cena za minutę rozmowy w ciągu dnia.

10. Obejrzyj alternatywny widok węzła *Linear Regression Learner*. Kliknij prawym przyciskiem myszy na węźle i wybierz opcję *View: Linear Regression Scatterplot View*.

11. Znajdź cenę minuty rozmowy nocnej i międzynarodowej.

2. Analiza skorelowanych zmiennych z wykorzystaniem regresji wielomianowej

1. Utwórz elementy przepływu do wczytania pliku *CarsWOO.txt*.

2. Przypomnij sobie, jak wygląda zależność między zmiennymi *mpg* i *weightlbs*. Obejrzyj te zmienne na wykresie *Scatter Plot*.

3. W celu zbadania zależności bliżej, wykorzystamy narzędzie regresji.

4. Przeanalizuj zależność pomiędzy zmiennymi *mpg* i *weightlbs*, wykorzystując regresję liniową.

5. Przeczytaj opis węzła *Polynomial Regression Learner* (folder w repozytorium węzłów *Analytics→Mining→Linear/Polynomial Regression*). Przeanalizuj zależność pomiędzy zmiennymi *mpg* i *weightlbs* jeszcze raz, wykorzystując regresję wielomianową stopnia 2.

3. Analiza zmiennych skorelowanych dla podzbiorów zbioru uczącego

1. Przenieś węzeł *Views→ Local (Swing)→Conditional Box Plot (local)* na przestrzeń roboczą projektu i połącz jego wejście z wyjściem węzła *File Reader*. Zapoznaj się z opisem węzła *Conditional Box Plot*.

2. Skonfiguruj węzeł *Conditional Box Plot*. Jako atrybut nominalny wybierz *brand*. Jako atrybut numeryczny wybierz *weightlbs*.

3. Uruchom węzeł za pomocą opcji *Execute and Open Views*. Obejrzyj wyniki. Przypomnij sobie, jak rozkłada się waga aut w zależności od miejsca pochodzenia.

4. Obejrzyj ponownie wykres tym razem dla atrybutu numerycznego *mpg*.

5. Podziel zbiór uczący na trzy części: samochody amerykańskie, japońskie i europejskie. (Wskazówka: można do tego wykorzystać węzły *Row filter*).

6. Dla każdej z części przeanalizuj zależność pomiędzy zmiennymi *mpg* i *weightlbs*, wykorzystując regresję liniową.

7. Dla każdej z części zbadaj siłę korelacji za pomocą węzła *Rank Correlation*.

II. Dyskretyzacja

Istnieją metody, które wymagają użycia konkretnego rodzaju atrybutu. Przykładowo regresja, k-średnie, PCA wymagają atrybutów numerycznych, natomiast algorytm Apriori atrybutów nominalnych.

Jako atrybut celu dla metod klasyfikacji również musimy podać atrybut nominalny.

4. Dyskretyzacja za pomocą przedziałów ustalonych przez analityka

1. Po uruchomieniu systemu i wybraniu odpowiedniej perspektywy (KNIME) przenieś na przestrzeń roboczą węzeł *File Reader*. Skonfiguruj go do odczytu pliku *cadata.csv*.

2. W pliku *cadata.csv* znajdują się dane opisujące domy w Kalifornii (dane pochodzą z 1990 roku z roczników statystycznych). Każdy przykład opisuje jeden kwartał miejski. Atrybuty to: mediana wartości domu (w dolarach), mediana wieku mieszkania, całkowita liczba pomieszczeń, całkowita liczba sypialni, liczba mieszkańców, liczba gospodarstw domowych, szerokość geograficzna i długość geograficzna.

3. Dołącz nowy węzeł *Interactive table* do wyjścia węzła *File Reader*. Uruchom węzeł *Interactive table* za pomocą opcji *Execute* and *Open Views* i obejrzyj dane.

4. Dodaj do przepływu węzeł *Manipulation→Column→Binning→Numeric Binner*. Podłącz jego wejście do wyjścia węzła File *Reader*. Zapoznaj się z opisem węzła *Numeric Binner*.

5. Skonfiguruj węzeł *Numeric Binner*:
    a. Przedziały będziemy ustalać dla atrybutu *median house value*. Wybierz ten atrybut z listy, a następnie ustaw dla niego trzy przedziały, trzykrotnie klikając przycisk *Add*.

   b. Ustal nazwy przedziałów kolejno na „domy niedrogie”, „domy drogie” i „rezydencje”. Granice przedziałów ustaw odpowiednio na 140000 i 300000.
   c. Upewnij się, że nowe wartości zastąpią dotychczasowe: pole *Append new column* musi pozostać niezaznaczone.

6. Dołącz nowy węzeł *Interactive table* do wyjścia węzła *Numeric Binner*. Uruchom węzeł *Interactive table* za pomocą opcji *Execute and Open Views* i zobacz efekt dyskretyzacji.

7. Dołącz nowy węzeł *Statistics* do wyjścia węzła *Numeric Binner*. Skonfiguruj węzeł tak, aby jako atrybut nominalny (w zielonej ramce) pozostał tylko *median house value*.

8. Uruchom węzeł *Statistics* za pomocą opcji *Execute and Open Views*. Przejdź na zakładkę dla atrybutów nominalnych i zapoznaj się z rozkładem wartości dla *median house value*.

5. Dyskretyzacja półautomatyczna

1. Dodaj do przepływu węzeł *Manipulation→Column→Binning→Auto Binner*. Podłącz jego wejście do wyjścia węzła *File Reader*. Zapoznaj się z opisem węzła *Auto Binner*.

2. Węzeł *Auto Binner* pozwala na utworzenie ustalonej liczby równych przedziałów lub na ustalenie przedziałów według kwantyli rozkładu. Skonfiguruj węzeł, wybierając tę drugą opcję:

   a. W ramce *Bining method* zaznacz pole *Sample quantiles*. Pozostaw domyślne kwantyle: „0.0, 0.25, 0.5, 0.75, 1.0”.
   b. W ramce *Bin naming* zaznacz *Borders*, dzięki czemu wartości dyskretyzowanego atrybutu będą opisywać konkretny przedział.
   c. Dyskretyzacji poddamy jedynie atrybut *median house value*. Usuń wszystkie atrybuty poza *median house value* z zielonej ramki *Include*.
   d. Chcemy, by nowe wartości zastąpiły dotychczasowe: zaznaczymy zatem pole *Replace target column(s)* u dołu okna.

3. Dołącz nowy węzeł *Interactive table* do wyjścia węzła *Auto Binner*. Uruchom węzeł *Interactive table* za pomocą opcji *Execute and Open Views* i zobacz efekt dyskretyzacji.

4. Dołącz nowy węzeł *Statistics* do wyjścia węzła *Auto Binner*. Skonfiguruj węzeł tak, aby jako atrybut nominalny (w zielonej ramce) pozostał tylko *median house value*.

5. Uruchom węzeł *Statistics* za pomocą opcji *Execute* and *Open Views*. Przejdź na zakładkę dla atrybutów nominalnych i zapoznaj się z rozkładem wartości dla *median house value*.