

Wstępne przygotowanie danych

Każdy ze studentów otrzyma rzeczywisty zbiór danych dotyczących kredytów 30-dniowych, czy dany klient powinien otrzymać kredyt. Do dyspozycji są następujące dane:

status pożyczki - atrybut określa, czy dany klient spłacił pożyczkę, spłacił częściowo, czy też wszczęty został proces windykacyjny wobec klienta.

kwota wnioskowana - kwota kredytu, o którą wnioskuje klient.

kwota otrzymana - kwota, która rzeczywiście otrzymał klient.

PESEL klienta - PESEL klienta, dla którego pozostawiono jedynie cyfry z których można odczytać płeć i datę urodzenia.

rodzaj źródła dochodu - forma prawna zatrudnienia, stanowiąca główne źródło dochodu klienta. Wyróżnia się tutaj następujące możliwości:

11 - umowa o pracę - czas nieokreślony;

12 - umowa o pracę - czas określony;

13 - emerytura;

14 - zasiłek przedemerytalny;

16 - renta stała;

17 - renta czasowa;

98 - inne - czas nieokreślony;

99 - inne - czas określony;

miesięczny dochód netto - miesięczny dochód netto na osobę w gospodarstwie domowym klienta.

które rolowanie - atrybut określa, które z rzędu rolowanie wykonał klient. Przez rolowanie rozumie się spłatę odsetek kredytu i zaciągnięcie kolejnego kredytu celem spłaty reszty zobowiązania nazywanych "chwilówkami". Celem laboratorium będzie przygotowanie danych do konstrukcji modelu decyzyjnego mającego na celu ocenę.

data decyzji - określa, kiedy została podjęta decyzja kredytowa.

data początkowa dochodu - określa od kiedy klient pobiera dochód z bieżącego źródła (od kiedy pracuje w bieżącym miejscu zatrudnienia, bądź od kiedy pobiera świadczenie emerytalne, rentę).

data końcowa dochodu - określa do kiedy klient będzie pobierał dochód z bieżącego źródła (w przypadku świadczeń na czas określony).

planowany termin spłaty - określa, kiedy klient powinien dokonać spłaty kredytu.

rzeczywisty termin spłaty - określa, kiedy klient dokonał spłaty kredytu (W przypadku spłaty częściowej (rolowania) określa, datę spłaty części zobowiązań, oraz zaciągnięcia nowego kredytu na resztę).

Ćwiczenie 1. W otrzymanym pliku .xls należy utworzyć kolumnę **kwota kredytu**. **Kwota kredytu** jest równa kwocie wnioskowanej, jeżeli w atrybucie **status pożyczki** wystąpiła odmowa, oraz jest równa kwocie otrzymanej w innym przypadku (należy wykorzystać funkcję **IF**)

Ćwiczenie 2. Należy utworzyć kolumnę **data urodzenia**. **Datę urodzenia** należy wyznaczyć z pierwszych 6 cyfr atrybutu **PESEL klienta** i dla każdego rekordu zapisać w postaci **YYYY-MM-DD** (sugeruje się wykorzystanie funkcji **CONCATENATE**, oraz **MID**, dla uproszczenia proszę przyjąć, że klienci urodzili się przed rokiem 2000)

Ćwiczenie 3. Należy utworzyć kolumnę **wiek** i wyznaczyć wiek klientów korzystając z atrybutów **data decyzji** i **data urodzenia** (można skorzystać z funkcji **YEAR**)

Ćwiczenie 4. Należy utworzyć kolumnę **pleć** i wypełnić odpowiednimi wartościami wykorzystując atrybut **PESEL klienta** (jeżeli 10 cyfra numeru PESEL jest parzysta, wówczas klient jest kobieta, w innym przypadku mężczyzna) (można korzystać z funkcji **MOD**)

Ćwiczenie 5. Należy utworzyć kolumnę **okres w jakim pobierał dochód** i wypełnić ją wartościami określającymi (w miesiącach) jak długo klient pobiera dochód na podstawie atrybutów **data początkowa dochodu**, oraz **data decyzji** (można skorzystać z funkcji **YEAR**, **MONTH**, **DAY**)

Ćwiczenie 6. Należy utworzyć kolumnę **okres w jakim będzie pobierał dochód** i wypełnić ją wartościami określającymi (w miesiącach) jak długo klient będzie pobierał dochód. W przypadku, gdy klient posiada umowę na czas nieokreślony, wówczas atrybut **data końcowa dochodu** będzie przyjmował wartość "0000-00-00 00:00:00". Zakładamy, że okres analizy obejmuje 4 lata, więc dla przypadków, dla których atrybut **data końcowa dochodu** będzie przyjmował wartość "0000-00-00 00:00:00", oraz dla okresów zatrudnienia dłuższych niż 48 miesięcy wartość atrybutu **okres w jakim będzie pobierał dochód** powinna być ustalona na 48.

Ćwiczenie 7. Należy utworzyć kolumnę **opóźnienie spłaty** i zapełnić ją wartościami określającymi (w dniach) jakie było opóźnienie w spłacie kredytu (biorąc pod uwagę

wartości atrybutów **planowany termin spłaty** i **rzeczywisty termin spłaty**). Jeżeli klient spłacił kredyt szybciej wówczas opóźnienie wynosi 0. Jeżeli klient nie otrzymał kredytu (atrybut **planowany termin spłaty** przyjmuje wartość "0000-00-00 00:00:00") wówczas przyjmujemy, że opóźnienie wyniesie 0. Jeżeli klient natomiast nie spłacił kredytu do tej pory (atrybut **rzeczywisty termin spłaty** przyjmuje wartość "0000-00-00 00:00:00") wówczas opóźnienie liczy się do dnia bieżącego.

Ćwiczenie 8. Utwórz plik XXXXXXL1 2.x/s zawierający jedynie kolumny (bez formuł tworzących) (należy skorzystać z opcji "Kopiuj tylko wartości"):

status pożyczki ;

kwota kredytu;

wiek;

płeć;

rodzaj źródła dochodu;

miesięczny dochód netto;

które rolowanie;

okres w jakim pobierał dochód;

okres w jakim będzie pobierał dochód;

opóźnienie spłaty.

Wykorzystując plik XXXXXXL1 2.x/s zbuduj odpowiadający mu plik danych w formacie .arff. Skonstruowany plik z danymi zapisz pod nazwa XXXXXXL2 2.arff. Zbadaj poprawność zbudowanego zbioru danych poprzez wczytanie go do interfejsu graficznego Weki.