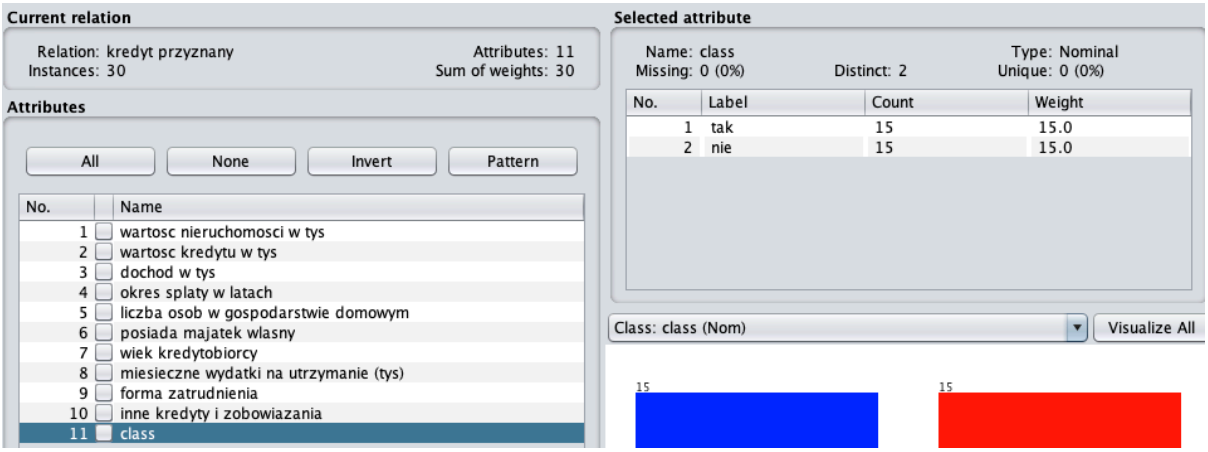
Zdolność kredytowa – zadanie z uczenia maszynowego (klasyfikacja)

Zdolność kredytowa wyznacza maksymalną kwotę kredytu, której bank może udzielić na dany cel (na przykład zakup mieszkania). Oprócz wysokości dochodów, wpływa na nią wiele innych czynników, takich jak miesięczny dochód, czy miesięczne wydatki na utrzymanie. W tym ćwiczeniu zostaną zbudowane modele pozwalające odpowiedzieć na pytanie, czy dany wniosek kredytowy powinien być przyjęty (tak), czy też odrzucony (nie) przez bank.

Zbiór danych

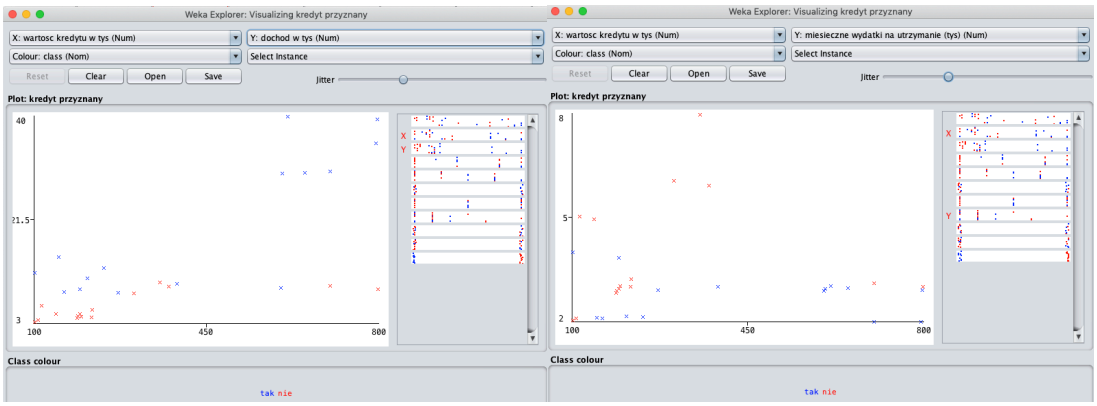
Przygotowany zbiór danych w formacie ARFF zawiera 30 obserwacji dla 10 atrybutów niezależnych (w tym trzy 3 kategoryczne i 7 numerycznych) oraz jeden atrybut zależny (kategoryczny). Zbiór jest doskonale zbalansowany (15 obserwacji na tak, 15 na nie).



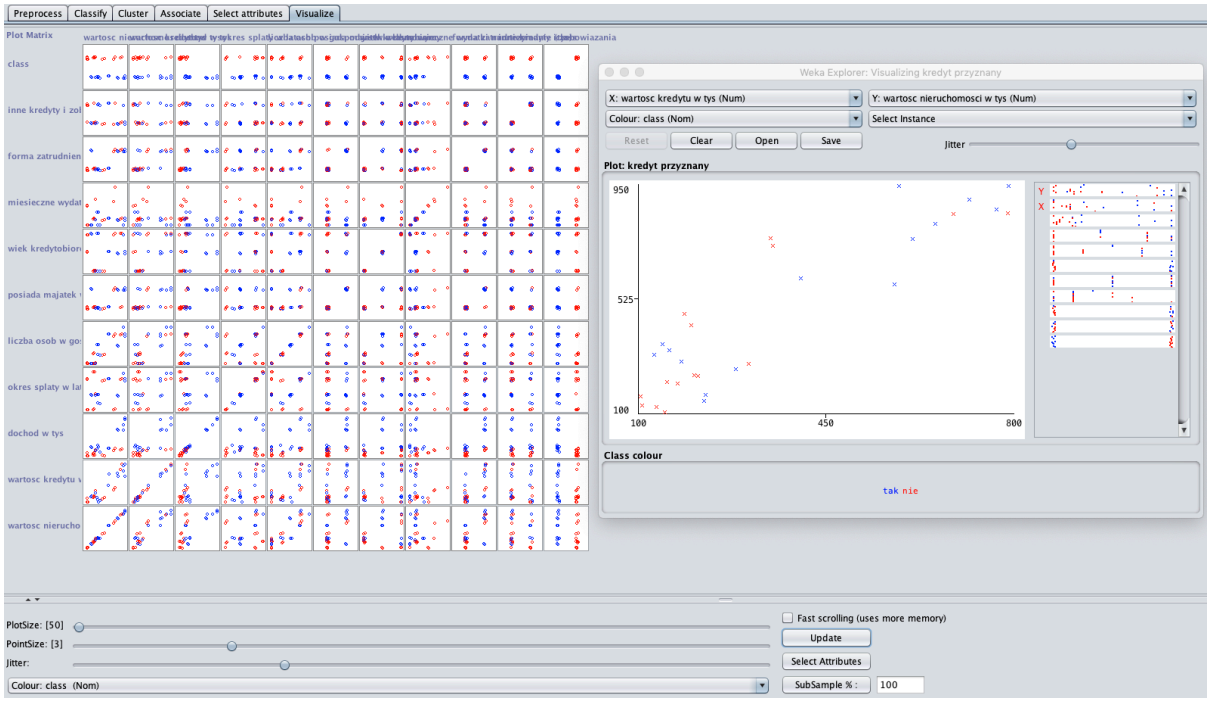
Celem ćwiczenia jest porównanie działania kilku klasyfikatorów i wybranie najlepszego z nich jako modelu decyzyjnego wspomagającego przydzielanie kredytów w przyszłości. Kryteriami oceny będą poprawność klasyfikacji oraz czytelność wiedzy.

Eksploracja danych

Pobieżna analiza wizualna pozwoliła na zidentyfikowanie silnych predyktorów najlepiej nadających się do klasyfikacji. Pierwszym z nich jest dochód. Im wyższy dochód, tym większa szansa na przyznanie kredytu. Dochód powyżej 12 tysięcy gwarantuje przyznanie kredytu. Z kolei miesięczne wydatki powyżej 5 tysięcy złotych oznaczają nieprzyznanie kredytu.



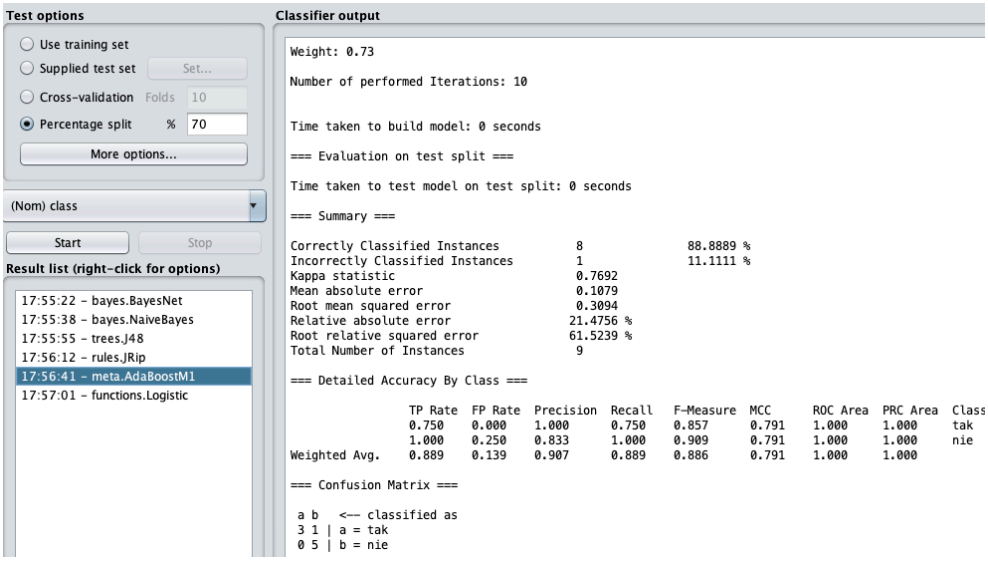
Ciekawą obserwacją jest też nie tyle zależność między wartością kredytu i ceną mieszkania (to nie jest zaskakujące), ale fakt, że większość odmów udzielenia kredytu miała miejsce dla kredytów o niższej wartości.



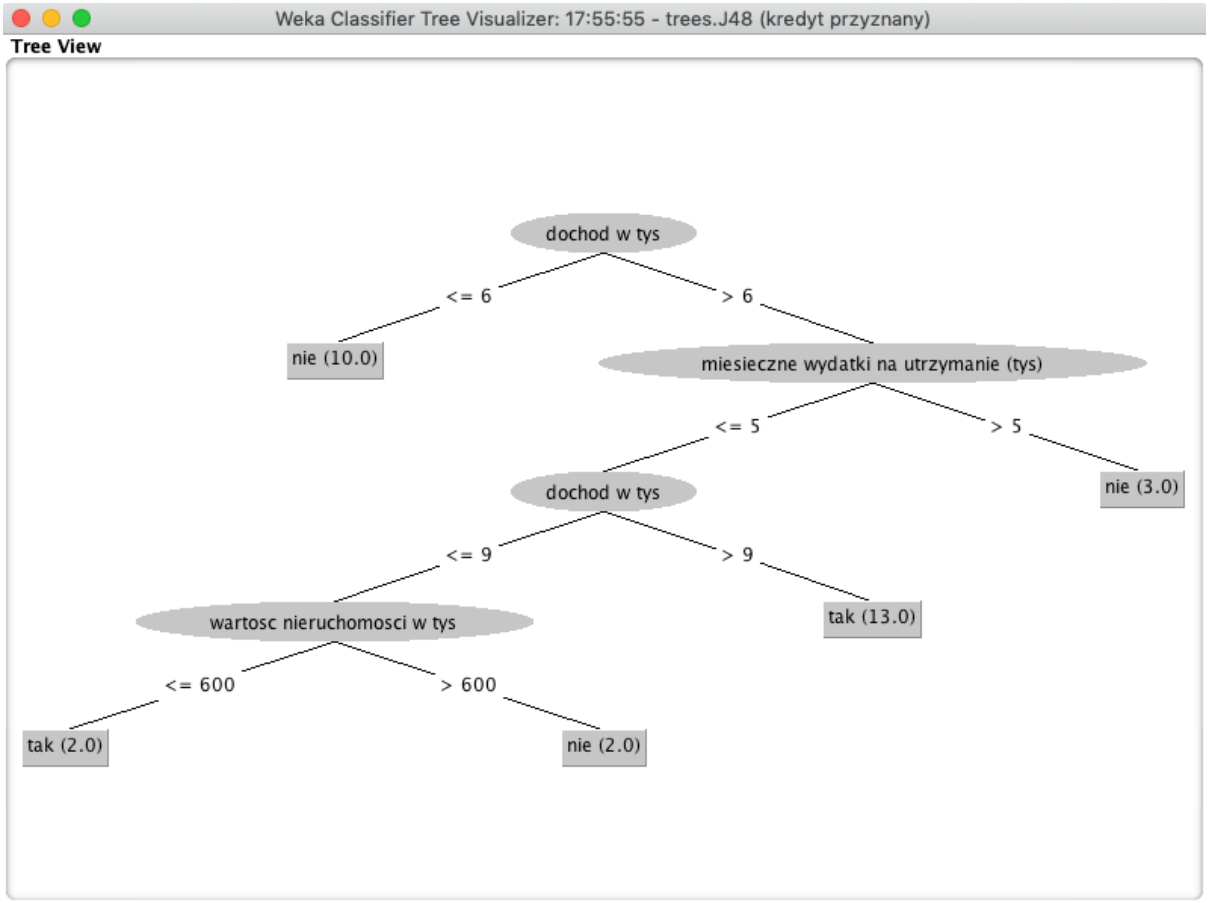
Modelowanie

Do wygenerowania klasyfikatorów wykorzystano moduł Explorer, algorytmy: BayesNet, NaiveBayes, J48, JRip, AdaBoostM1, Logistic oraz technikę testowania opartą na podziale na zbiór uczący (70%, 21 obserwacji) i testowy (30%, 9 obserwacji).

Najlepiej z zadaniem poradził sobie algorytm AdaBoostM1, uzyskując F-Measure na poziomie 0.886 przy 89% (8) poprawnie sklasyfikowanych oraz tylko 11% (1) niepoprawnie sklasyfikowanych obserwacji.



Algorytm J48 (F-Measure 0.778) wygenerował drzewo decyzyjne (poniżej), z którego wynika, iż nie ma co marzyć o kredycie, jeśli zarabia się mniej niż 6 (tys.) oraz miesięczne wydatki ma się na poziomie powyżej 5 (tys.). W przypadku wartości nieruchomości powyżej 600 (tys.) trzeba zarabiać powyżej 9 (tys.) aby otrzymać kredyt.



Podobne informacje otrzymamy z JRip (F-Measure 0.778), chociaż tutaj próg akceptowalnych miesięcznych wydatków został podniesiony do 6 (tys.), a przy wartościach kredytu powyżej 700 (tys.) i wartości nieruchomości poniżej 850 (tys.) również otrzymamy odmowę.

```
JRIP rules:
===========

(dochod w tys <= 6) => class=nie (10.0/0.0)
(miesieczne wydatki na utrzymanie (tys) >= 6) => class=nie (3.0/0.0)
(wartosc kredytu w tys >= 700) and (wartosc nieruchomosci w tys <= 850) => class=nie (2.0/0.0)
 => class=tak (15.0/0.0)

Number of Rules : 4
```

Na koniec użyto modułu Experimenter do uzyskania wyników porównawczych. Metoda podziału na zbiór uczący oraz testowy była taka sama, wybór obserwacji do poszczególnych zbiorów losowy (stad nieco inne wyniki, chociaż co do ostatecznej konkluzji zgodne). Klasyfikator oparty o algorytm AdaBoostM1 okazał się najlepszy oraz statystycznie (poziom

istotności 5%) lepszy od referencyjnego klasyfikatora zbudowanego za pomocą algorytmu BayesNet.

Test output

```
Tester:     weka.experiment.PairedTTester -G 3,4,5 -D 1 -R 2 -S 0.05 -result-matrix "we
Analysing:  F_measure
Datasets:   1
Resultsets: 6
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       15/03/2020, 18:26


Dataset                   (1) bayes.B | (2) baye (3) tree (4) rule (5) meta (6) func
----------------------------------------------------------------------------------
'kredyt przyznany'        (10)   0.70 |   0.80     0.76     0.71     0.80 v   0.77
----------------------------------------------------------------------------------
                              (v/ /*) |  (0/1/0)  (0/1/0)  (0/1/0)  (1/0/0)  (0/1/0)


Key:
(1) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estim
(2) bayes.NaiveBayes '' 5995231201785697655
(3) trees.J48 '-C 0.25 -M 2' -217733168393644444
(4) rules.JRip '-F 3 -N 2.0 -O 2 -S 1' -6589312996832147161
(5) meta.AdaBoostM1 '-P 100 -S 1 -I 10 -W trees.DecisionStump' -1178107808933117974
(6) functions.Logistic '-R 1.0E-8 -M -1 -num-decimal-places 4' 3932117032546553727
```

Z punktu widzenia czytelności wiedzy (w tym przypadku bardzo istotnej, gdyż bank musi być w stanie uzasadnić decyzję odmowną), algorytmy J48 lub JRip byłyby bardziej preferowane.

Wynik

Ostatecznie J48 osiągnął najlepszy wynik w ogólnym porównaniu (poprawność klasyfikacji oraz czytelność wiedzy).