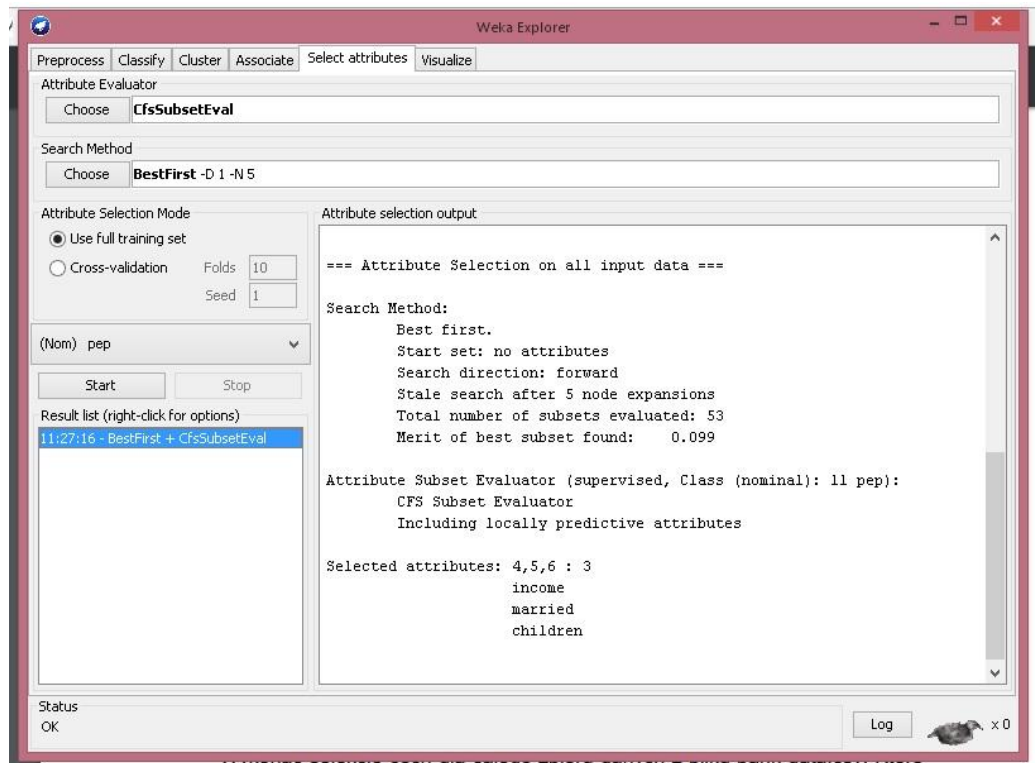


Edyta Karasek

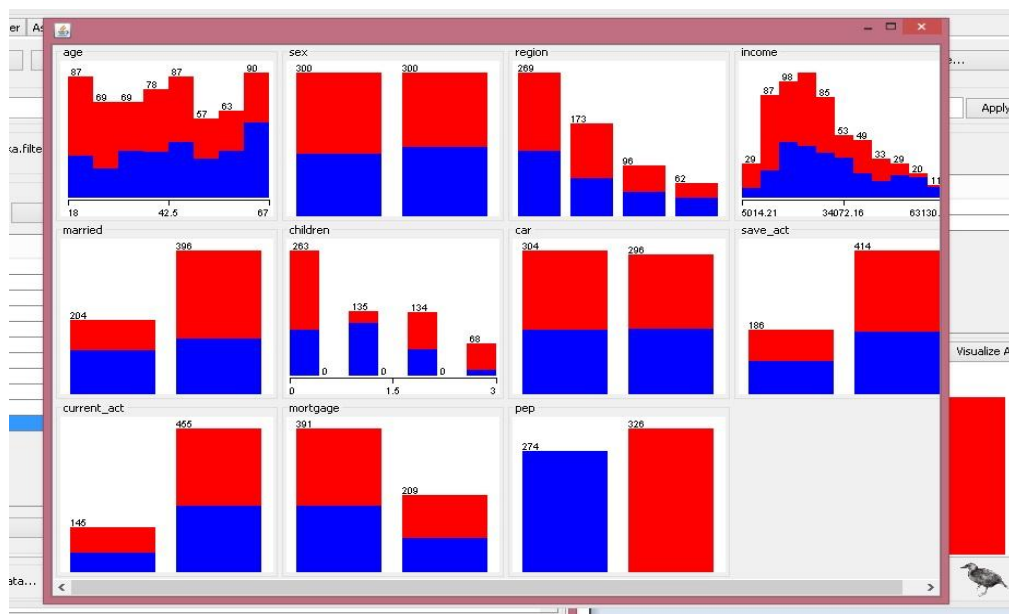
IWB, Studia niestacjonarne, 1 rok, 2 stopień

Zadanie 1. Selekcja cech (Weka / Explorer)

Po usunięciu atrybutu id, została dokonana selekcja. Metodą BestFirst zostały wybrane 3 atrybuty: income, married i children.

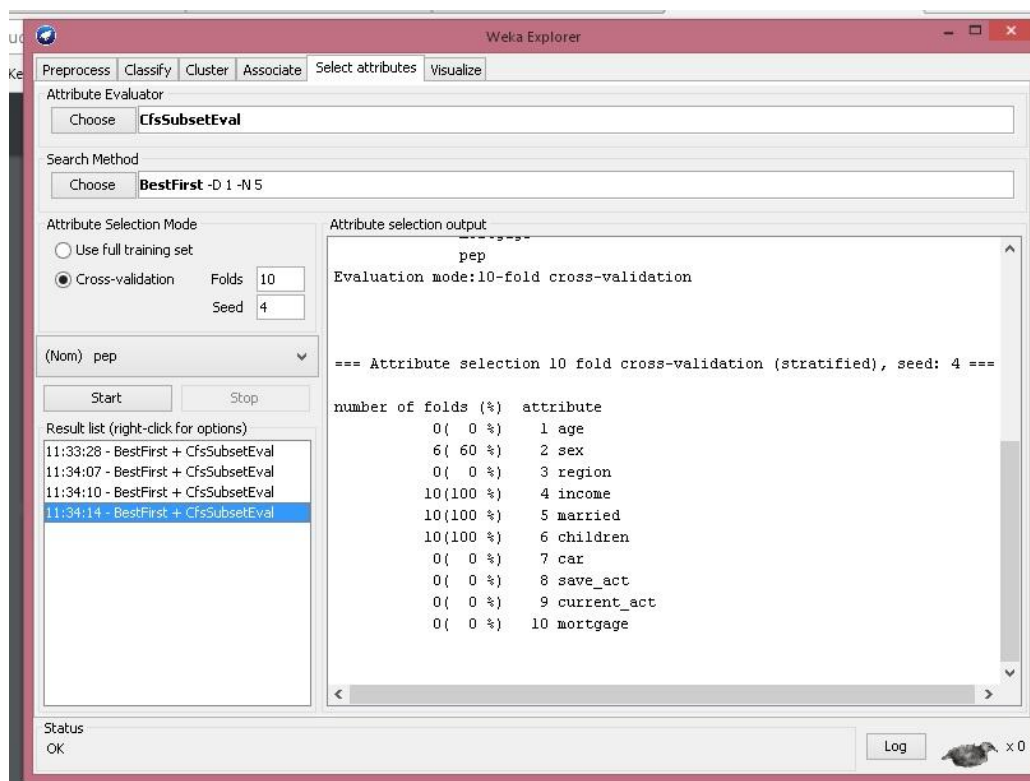


Biorąc pod uwagę ocenę wizualną można wskazać również te same atrybuty, ponieważ charakteryzują się one największą różnorodnością, co jest równoznaczne z przekazaniem największej ilości informacji na temat preferencji klientów.

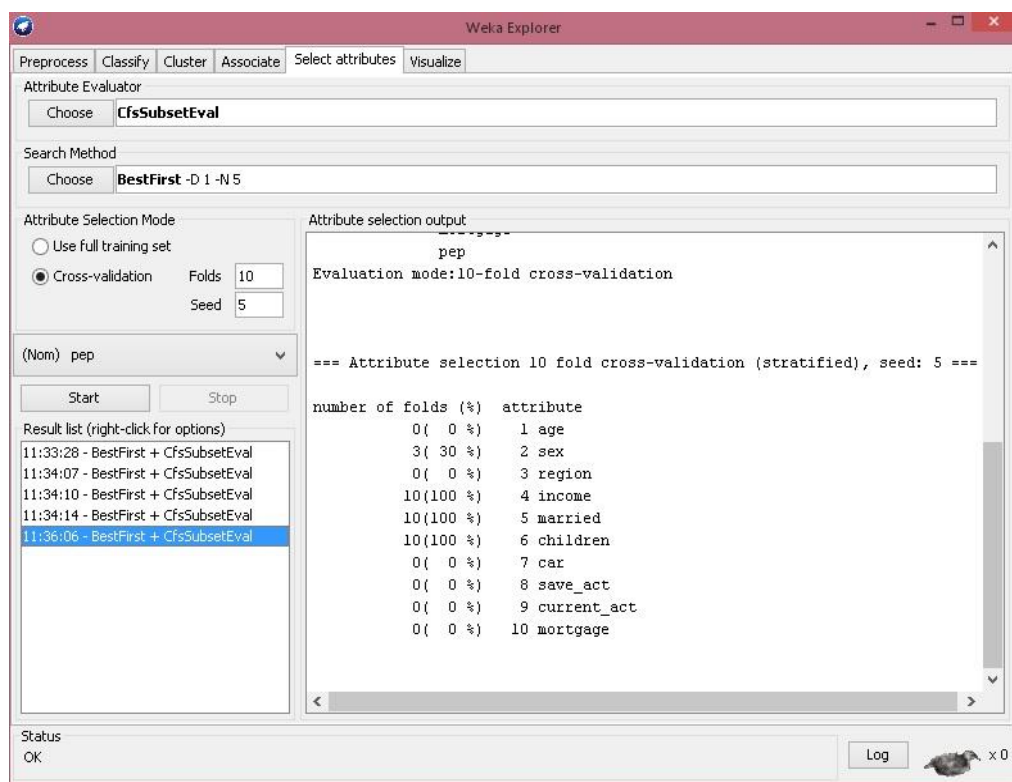


Następnie przy użyciu metody walidacji skróśnej modelując i zmieniając wartość seed wytypowano dodatkowe atrybuty, które niosą użyteczne informacje na temat preferencji klientów.

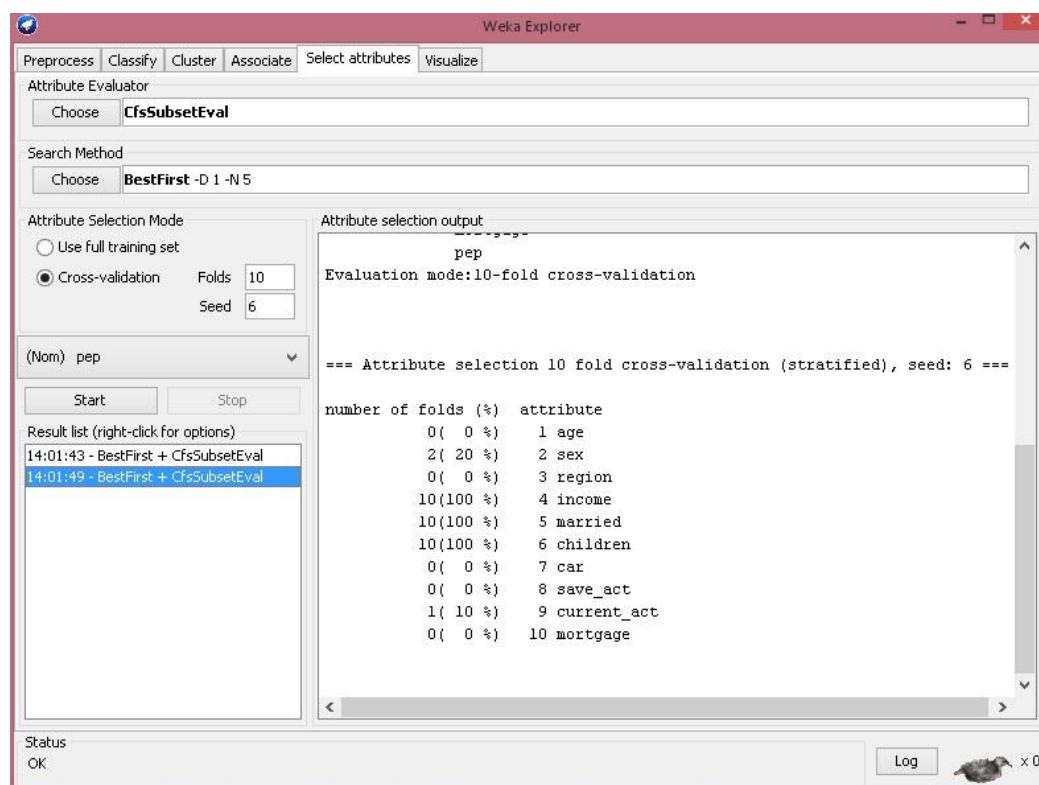
Przy określeniu wartości Seed równej 4 dokonano wyboru atrybutu dotyczącego płci (sex) klientów.



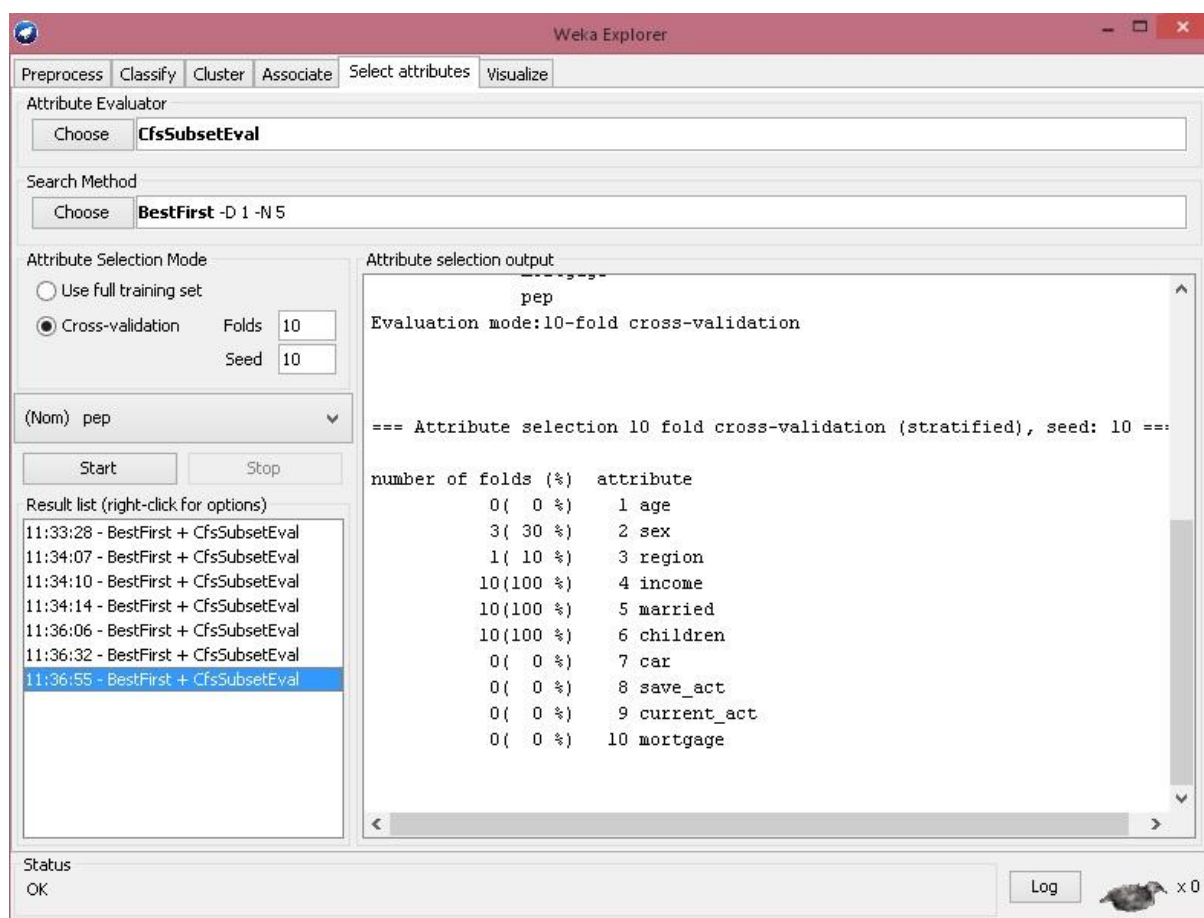
Przy ustaleniu wartości seed równej 5 nie wykryto żadnego potencjalnie nowego atrybutu.



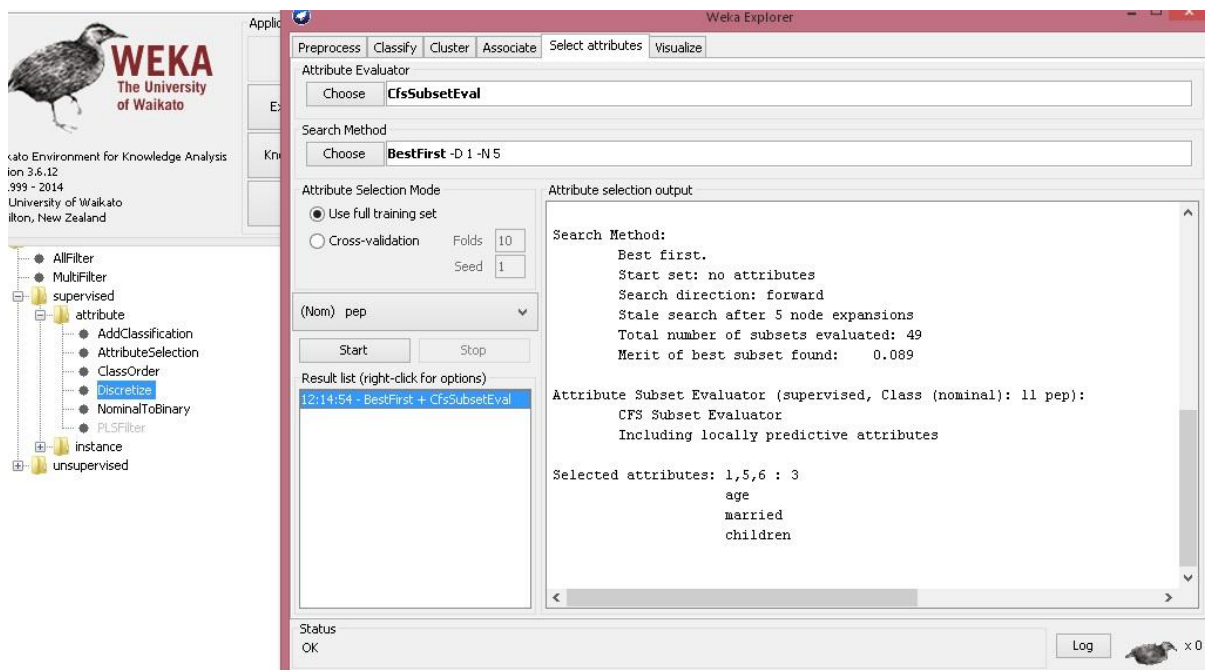
Przy ustaleniu wartości Seed równej 6 wykryto dodatkowo atrybut „current_act”.



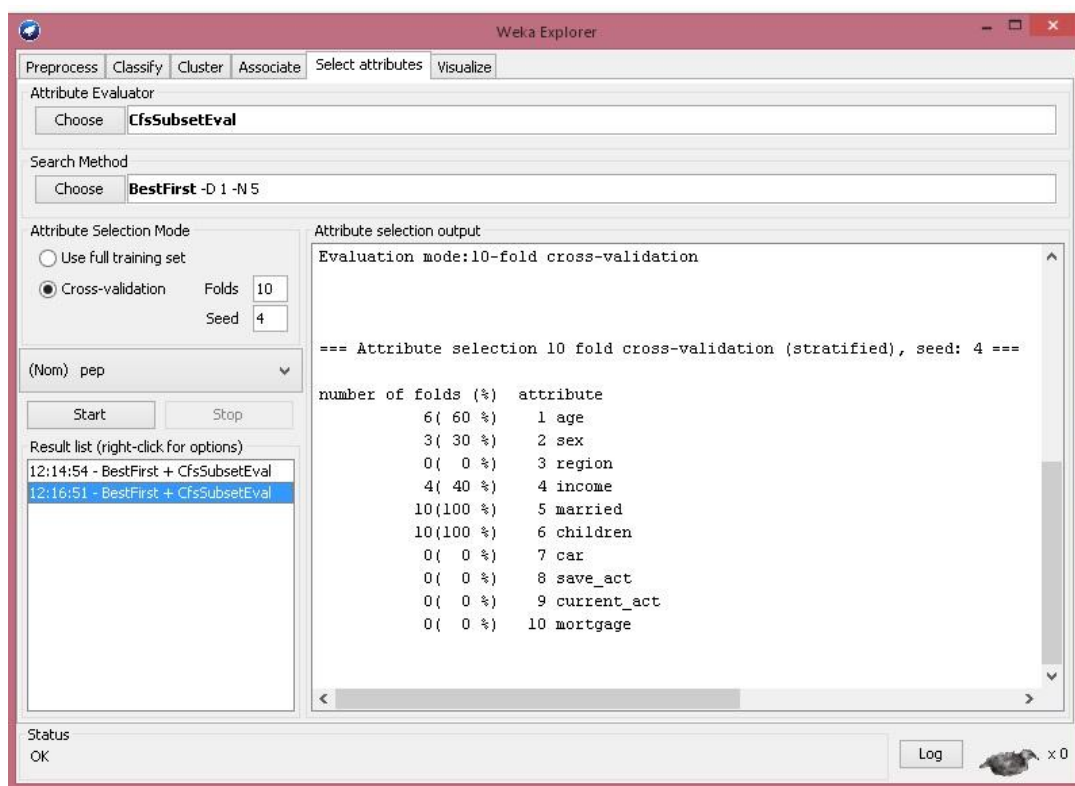
Przy ustaleniu wartości Seed równej 10 do grona atrybutów dołączył region.

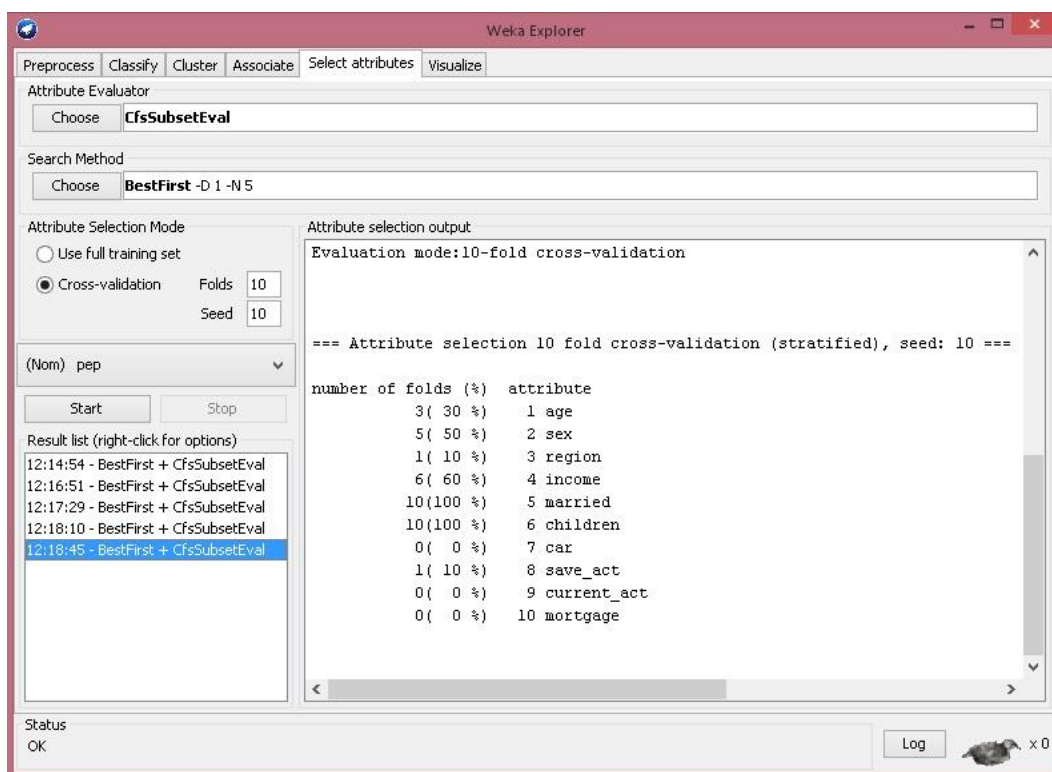
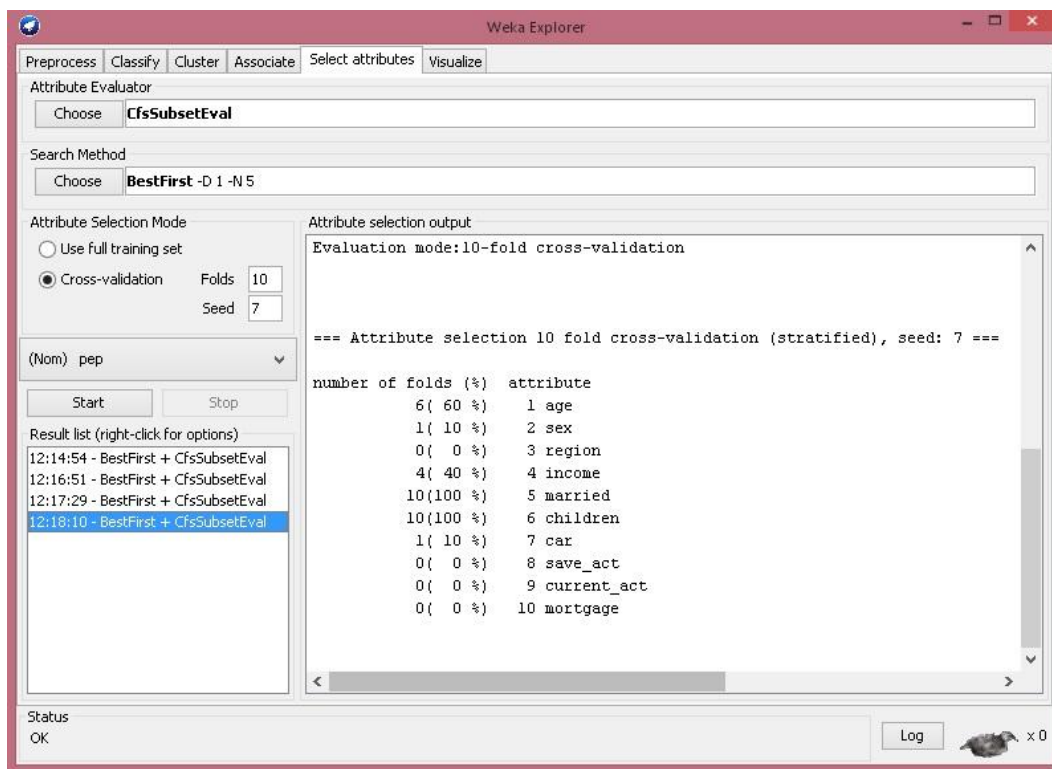


Następnie zgodnie z poleceniem dokonano dyskretyzacji parametru income na 2 przedziały. Do zbioru wyselekcjonowanych atrybutów trafiły „age”, „married” i „children”, nie trafił natomiast atrybut „income”.



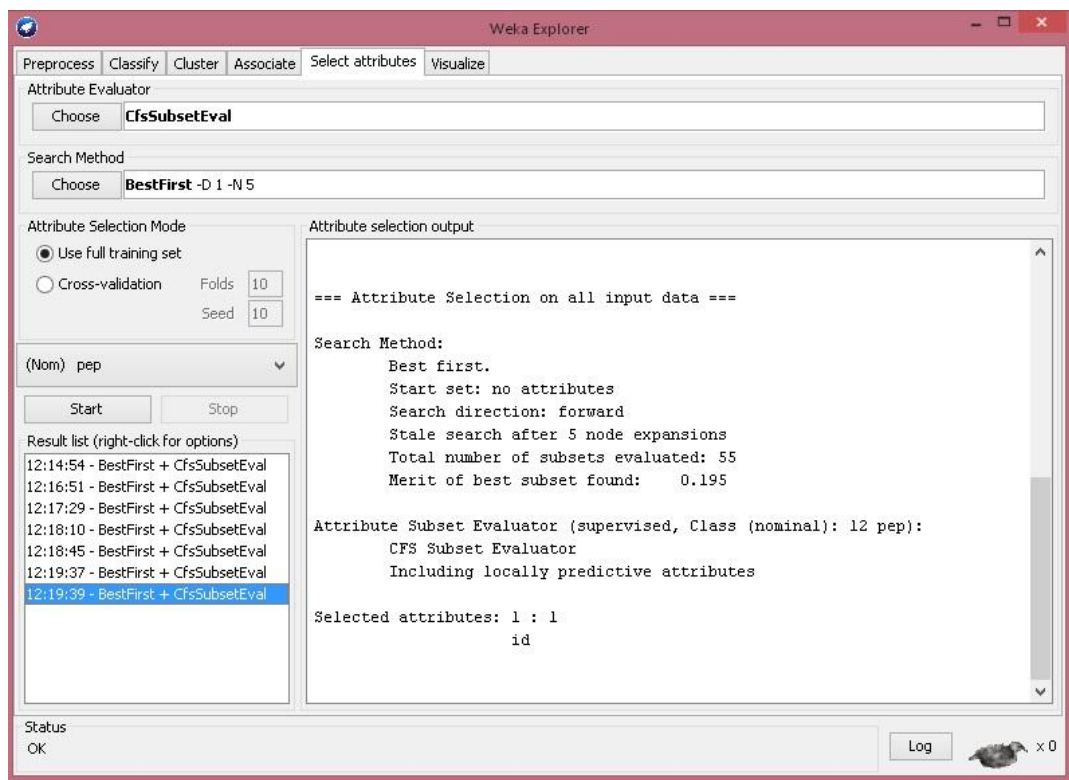
Kolejnym punktem było wykonanie walidacji skrośnej. Dokonano ponownie obserwacji „przydatności” atrybutu „income” w zależności od wartości parametru seed.





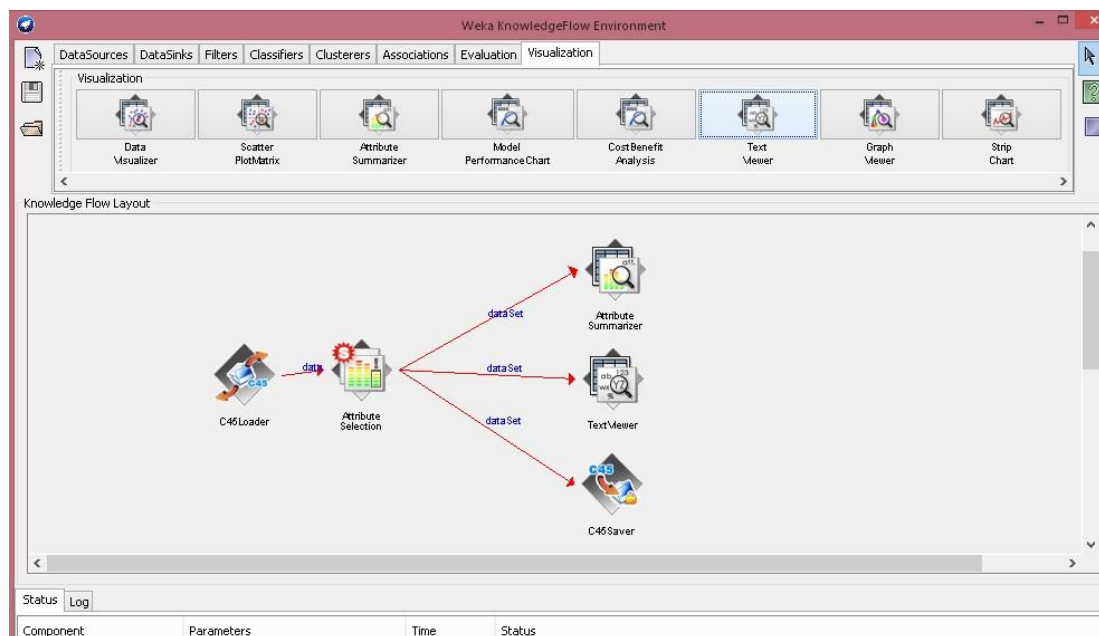
Po dyskretyzacji atrybutu „income” można zauważyć, że jego przydatność do klasyfikacji klientów wzrosła.

Ostatnim punktem pierwszego zadania, było sprawdzenie jak będzie przebiegać selekcja cech, w momencie gdy zachowamy atrybut id, pomijając jego usunięcie. Na podstawie wyników zaobserwowano, że przy pozostawieniu tego atrybutu w zbiorze, jest on jedynym istotnym parametrem. W rzeczywistości jednak nie ma żadnej użyteczności z pozostawienia go w bazie.



Zadanie 2. Selekcja cech (Weka / KnowledgeFlow)

Początkowo został utworzony przepływ, który realizuje selekcję cech.

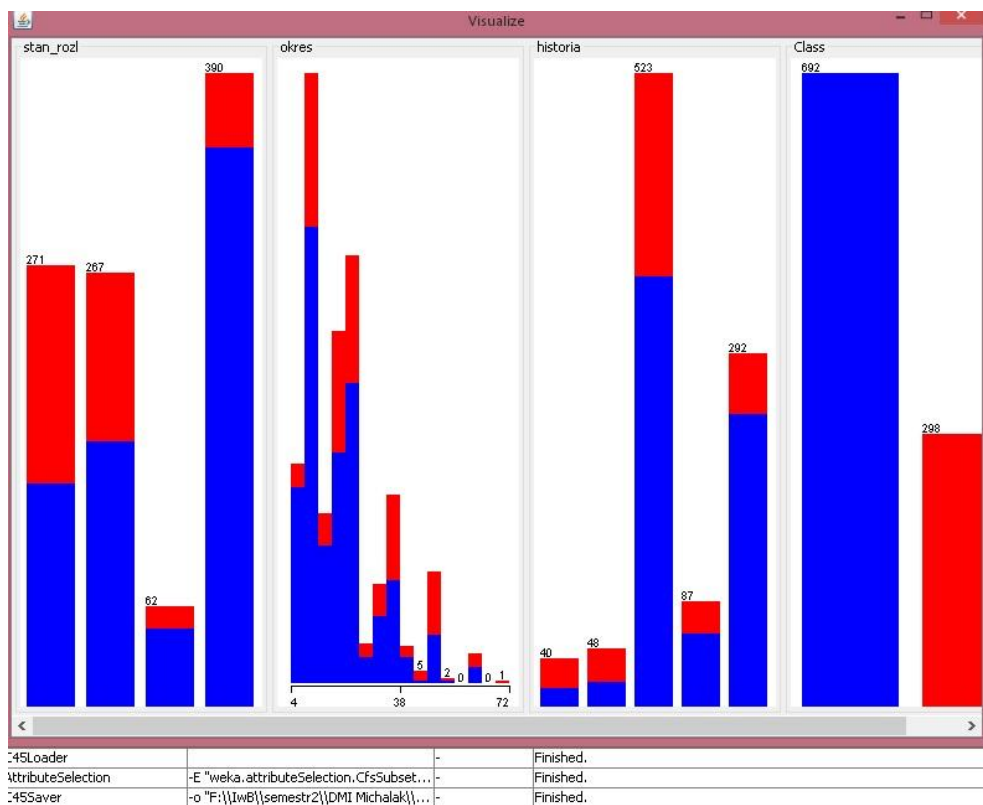


Następnie został przypisany plik wejściowy, z którego zostały pobrane i załadowane dane do przepływu.

Kolejnym punktem było utworzeniu i wskazanie katalogu wyjściowego do przepływu. Po załadowaniu danych wygenerowane pliki zostały automatycznie umieszczone w wybranym folderze.

Ostatnim punktem było porównanie wyników selekcji atrybutów dla przeszukiwania (search) metodą BestFirst i GeneticSearch.

Metoda BestFirst:



Text Viewer

Result list

12:36:23 - kredyty_20-weka

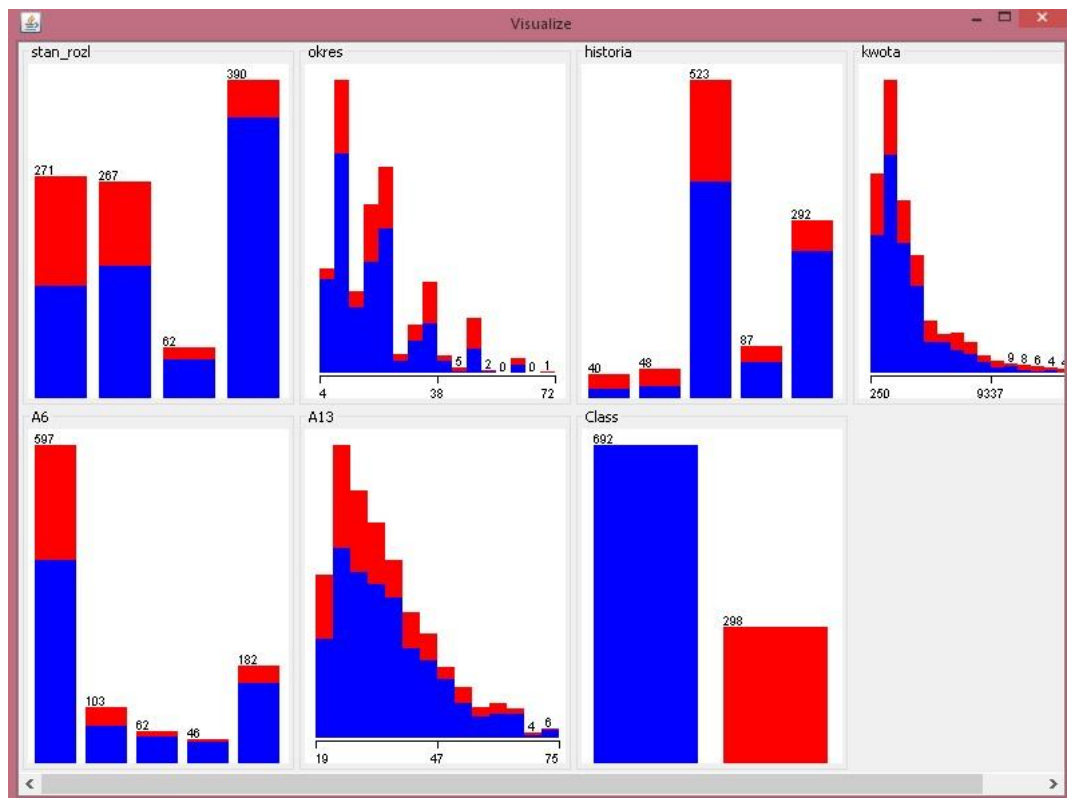
Text

```
@relation 'kredyty_20-weka.filters.supervised.attribute.AttributeSelection-Ewe'

@attribute stan_rozl {dlug,malo,duzo,brak}
@attribute okres numeric
@attribute historia {wszystkie_ok,w_banku_ok,biezace_ok,mial_spoznienia,krytyc
@attribute Class {dobry,zly}

@data
malo,12,biezace_ok,zly
dlug,48,biezace_ok,zly
malo,12,biezace_ok,dobry
dlug,24,krytyczny,zly
dlug,24,biezace_ok,zly
brak,24,krytyczny,dobry
dlug,30,wszystkie_ok,dobry
malo,24,biezace_ok,zly
brak,24,biezace_ok,dobry
brak,9,krytyczny,dobry
dlug,6,biezace_ok,dobry
dlug,10,krytyczny,dobry
malo,12,krytyczny,dobry
brak,10,krytyczny,dobry
dlug,6,biezace_ok,dobry
brak,6,wszystkie_ok,dobry
malo,7,biezace_ok,dobry
dlug,60,mial_spoznienia,zly
malo,18,biezace_ok,dobry
dlug,24,biezace_ok,dobry
malo,18,biezace_ok,dobry
brak,12,krytyczny,dobry
duzo,12,biezace_ok,dobry
malo,45,krytyczny,zly
brak,48,krytyczny,dobry
duzo,18,biezace_ok,zly
duzo,10,biezace ok.dobry
```

Metoda GeneticSearch



Result list

12:36:23 - kredyty_20-weka.filters
12:40:53 - kredyty_20-weka.filters

Text

```
@relation 'kredyty_20-weka.filters.supervised.attribute.AttributeSelection-Eweka.attrib

@attribute stan_rozl {dlug,malo,duzo,brak}
@attribute okres numeric
@attribute historia {wszystkie_ok,w_banku_ok,biezace_ok,mial_spoznienia,krytyczny}
@attribute kwota numeric
@attribute A6 {A61,A62,A63,A64,A65}
@attribute A13 numeric
@attribute Class {dobry,zly}

@data
malo,12,biezace_ok,1295,A61,25,zly
dlug,48,biezace_ok,4308,A61,24,zly
malo,12,biezace_ok,1567,A61,22,dobry
dlug,24,krytyczny,1199,A61,60,zly
dlug,24,biezace_ok,1282,A62,32,zly
brak,24,krytyczny,2424,A65,53,dobry
dlug,30,wszystkie_ok,8072,A65,25,dobry
malo,24,biezace_ok,12579,A61,44,zly
brak,24,biezace_ok,3430,A63,31,dobry
brak,9,krytyczny,2134,A61,48,dobry
dlug,6,biezace_ok,2647,A63,44,dobry
dlug,10,krytyczny,2241,A61,48,dobry
malo,12,krytyczny,1804,A62,44,dobry
brak,10,krytyczny,2069,A65,26,dobry
dlug,6,biezace_ok,1374,A61,36,dobry
brak,6,wszystkie ok,426,A61,39,dobry
```

00:22:40 - Now saved.
- Finished.
-E "weka.attributeSelection.CfsSubset... - Finished.
-o "F:\IwB\semestr2\DMI Michalak\... - Finished.

W zależności od metody zostały wytypowane różne atrybuty. Odnosząc się do selekcji typu Genetic Search lista wybranych cech jest większa. Na podstawie metody BestFirst wytypowane zostały :stan_rozl, okres i historia. Selekcja GeneticSearch zawierała atrybuty z metody porównywalnej oraz dodatkowo atrybuty taki jak kwota, A6 i A13.