Monika Szymańska 238359

Zadanie 3.3

3.3.1 grupowanie w aplikacji RapidMiner

Do nowego procesu stworzonego w aplikacji RapidMiner został wczytany plik csv zawierający dane o klientach. Przy wczytywaniu danych zostały mienione typy pól zawierające numer klienta oraz miesiąc z numeric na polynominal, pole przedzialkwotowy oznaczono jako zmienną wyjściową, a numerklienta jako id. Po dodaniu do modelu bloków k-Means oraz Extract Cluster Prototypes został uruchomiony proces a jego wyniki prezentują się w następujący sposób:

Cluster Model

```
Cluster 0: 218 items
Cluster 1: 0 items
Cluster 2: 207 items
Cluster 3: 327 items
Cluster 4: 248 items
Total number of items: 1000
```

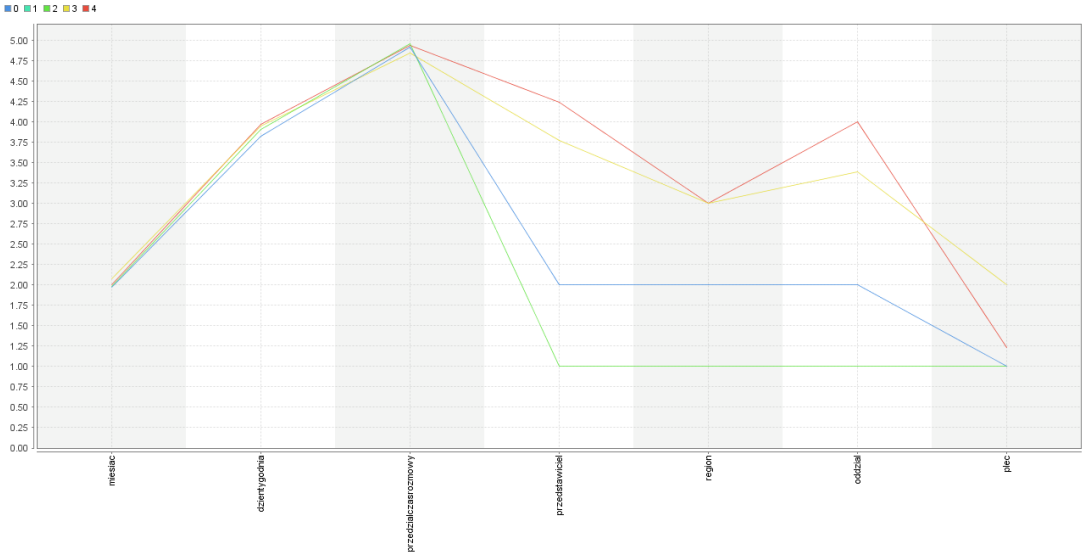
Rysunek 1 - Wyniki grupowania danych o klientach metodą k-Means

Jak można zaobserwować na Rysunku 1 dane o klientach zostały podzielone na 5 grup, z których jedna jest pusta. Najbardziej liczbą grupę bo zawierającą aż 327 argumentów zawiera Cluster 3, nieco mniej bo 248 argumentów zostało przyporządkowanych do grupy ostatniej, natomiast w Cluster 0 i 2 znalazło się kolejno 218 oraz 207 obiektów. Centra powstałych skupień zostały zaprezentowane w Tabeli 1.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
miesiac	1.968	?	1.986	2.080	2
dzientygodnia	3.830	?	3.908	3.954	3.968
przedzialczasrozmowy	4.922	?	4.957	4.847	4.944
przedstawiciel	2	?	1	3.777	4.238
region	2	?	1	3	3
oddzial	2	?	1	3.388	4
plec	1	?	1	2	1.238

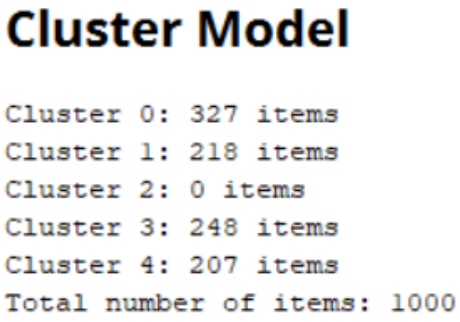
Tabela 1 - Centra skupień dla zgrupowanych danych o klientach

Jak widać najbardziej rozbieżne wartości możemy odnotować dla pól przedstawiciel, region i oddział (nieco mniej dla zmiennej płeć) co doskonale widać na poniższym wykresie. Można zatem stwierdzić, że to te zmienne były najbardziej kluczowymi podczas klasyfikacji.



Rysunek 2 - Wykres przedstawiający wartości centrów skupień dla zgrupowanych danych o klientach metodą k-Means

Kolejnym etapem było dodanie nowego strumienia, będącego kopią poprzedniego bez zmiennej numerklienta. Otrzymany proces został zapisany pod nazwą proces_3.3_2, a jego wyniki prezentuje Rysunek 3.



Rysunek 3 - Wyniki grupowania danych o klientach z pominięciem zmiennej numerklienta metodą k-Means

Jak możemy zauważyć wielkości stworzonych grup są takie same. Analizując centra powstałych skupień (Tabela 2) również widzimy, że niczym się one nie różnią.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
miesiac	2.080	1.968	?	2	1.986
dzientygodnia	3.954	3.830	?	3.968	3.908
przedzialczasrozmowy	4.847	4.922	?	4.944	4.957
przedstawiciel	3.777	2	?	4.238	1
region	3	2	?	3	1
oddzial	3.388	2	?	4	1
plec	2	1	?	1.238	1

Tabela 2 - Centra skupień dla zgrupowanych danych o klientach z pominięciem zmiennej numerklienta

Następnie został utworzony kolejny strumień, w którym zostały wczytane dane z pliku klienci3. Korzystając z pola zawierającego datę, stworzono dwa kolejne z miesiącem i dniem tygodnia oraz wykorzystano operatory Aggreagte do zsumowania czasu rozmowy oraz kwoty dla klientów oraz przedstawicieli. Wyniki dokonanych agregacji dla klientów przedstawione ostały w Tabeli 3.

Row No.	numer klienta	sum(kwota ...	sum(czas ro...
1	1	4751	3768
2	10	9025	6674
3	11	10607	8742
4	12	5791	5261
5	13	4489	4014
6	14	5632	4511
7	15	7364	4721
8	2	7761	5756
9	3	8917	6029
10	4	11529	8576
11	5	9931	7473
12	6	10377	7117
13	7	9763	7605
14	8	8916	6021
15	9	8031	6302

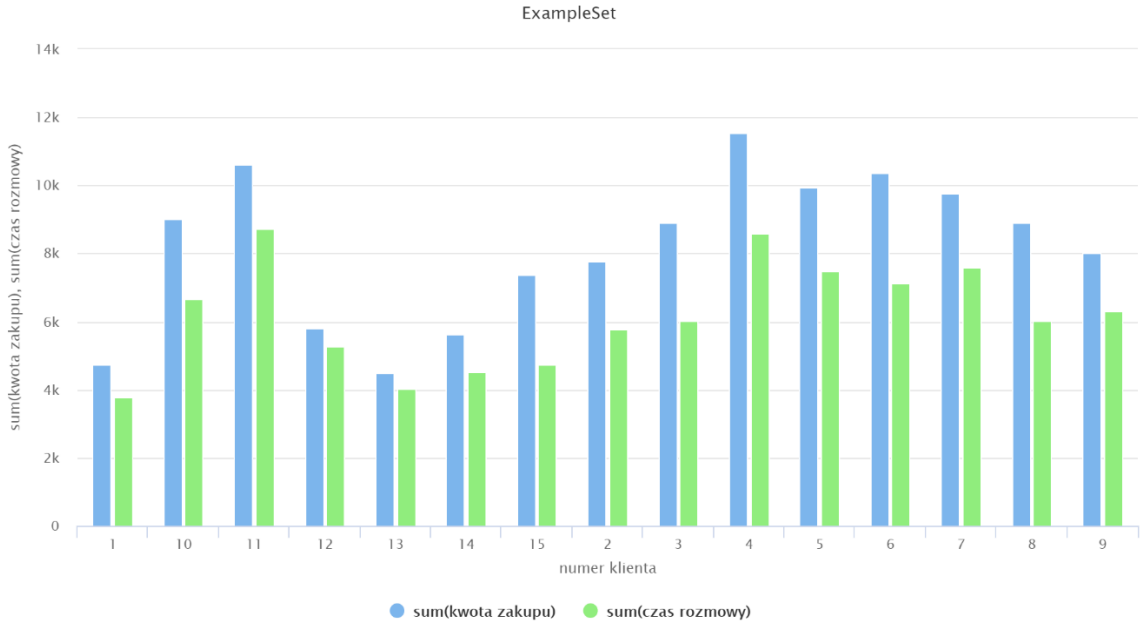
Tabela 3 - Wyniki agregacji kwot zakupu oraz czasów rozmowy dla klientów

Wyniki agregacji dla przedstawicieli prezentuje Tabela 4.

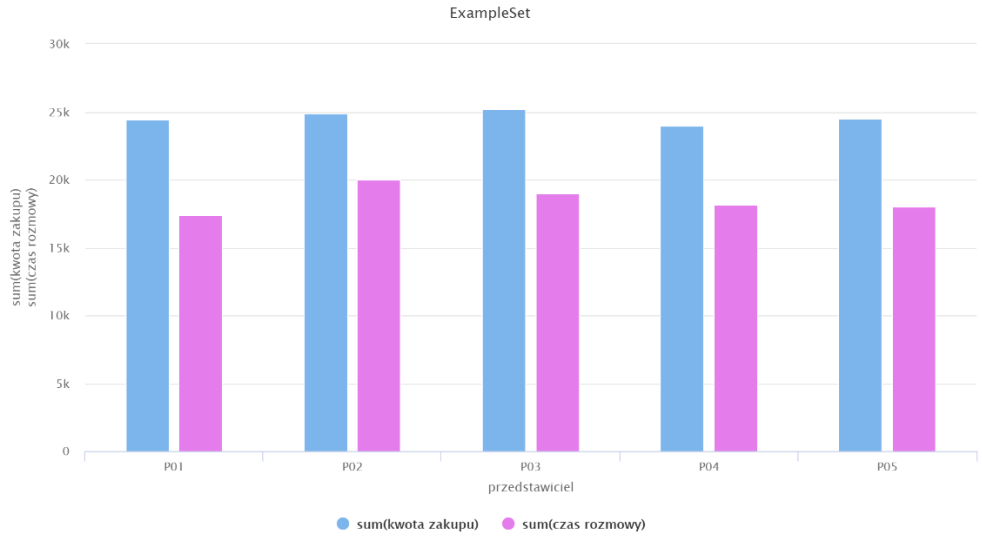
Row No.	przedstawic...	sum(czas ro...	sum(kwota ...
1	P01	17388	24426
2	P02	19976	24857
3	P03	19022	25155
4	P04	18192	23976
5	P05	17992	24470

Tabela 4 - Wyniki agregacji kwot zakupu oraz czasów rozmowy dla przedstawicieli

Dodatkowo zakładka Visualization pozwala nam obejrzeć otrzymane wyniki na wykresach. Poniżej zostały przedstawione przykładowe wizualizacje otrzymane w wyniku zsumowania kwot zakupu i czasów rozmów dla klientów oraz przedstawicieli.



Rysunek 4 - Wykres przedstawiający sumaryczne kwoty zakupu i czasy rozmów dla klientów



Rysunek 5 - Wykres przedstawiający sumaryczne kwoty zakupu i czasy rozmów dla przedstawicieli

Do zagregowanych danych dołączono operatory k-Menas oraz Extract Cluster Prototypes, a otrzymany proces zapisano pod nazwą proces3.3_4. Wynik grupowania dla klientów został przedstawiony na Rysunku 6.

Cluster Model

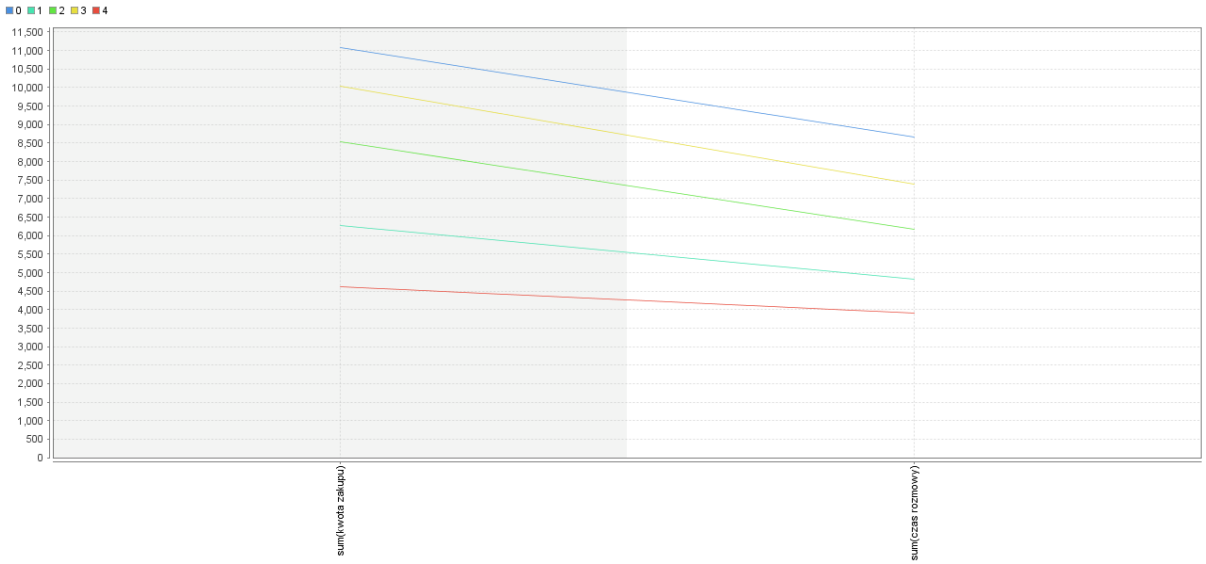
```
Cluster 0: 2 items
Cluster 1: 3 items
Cluster 2: 5 items
Cluster 3: 3 items
Cluster 4: 2 items
Total number of items: 15
```

Rysunek 6 - Wynik grupowania dla klientów ze zbioru danych klienci3

Jak widać 15 klientów zostało podzielonych na 5 grup z których najbardziej liczna jest grupa trzecia – Cluster 2, ponieważ zawiera aż 5 klientów. Centra powstałych skupień prezentują się natomiast następująco:

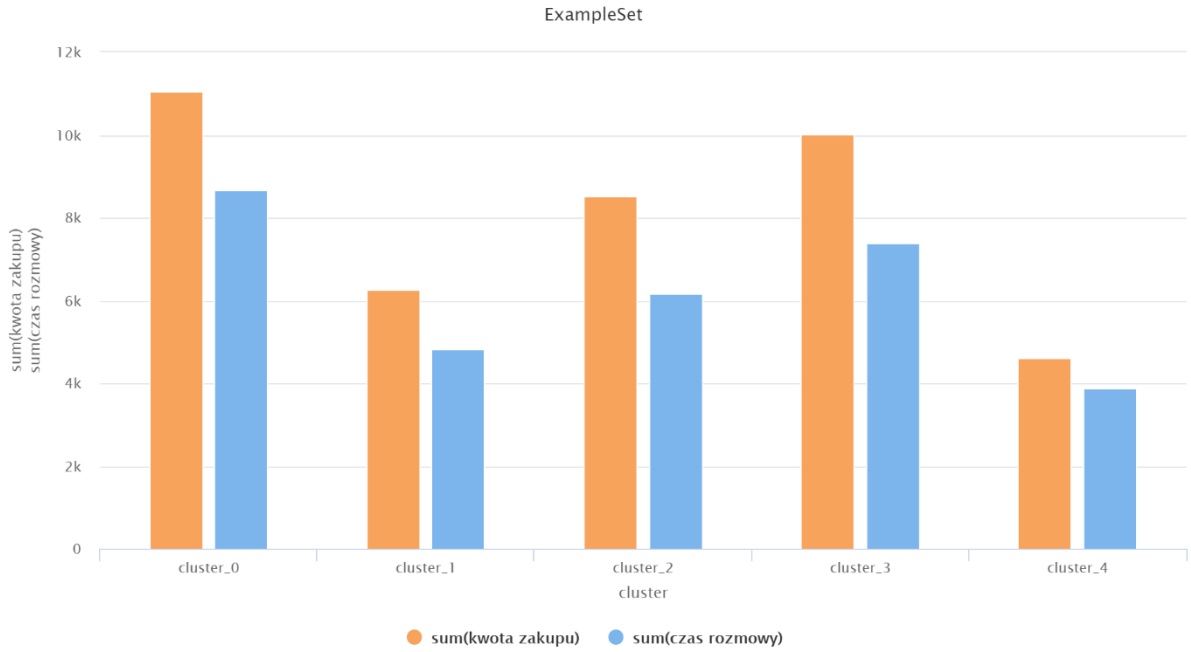
Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
sum(kwota zakupu)	11068	6262.333	8530	10023.667	4620
sum(czas rozmowy)	8659	4831	6156.400	7398.333	3891

Tabela 5 - Centra skupień dla zgrupowanych metodą k-Means klientów ze zbioru danych klienci3



Rysunek 7 - Centra skupień dla zgrupowanych metodą k-Means klientów ze zbioru danych klienci3

Jak widać na Rysunku 7 oraz w Tabeli 5 Cluster 0 zawiera klientów o najwyższych sumarycznych kwotach zakupu oraz najdłuższych sumarycznych czasach rozmów, natomiast dla kolejnych grup te wartości są coraz niższe. Najwięcej klientów zawierał Cluster 2, a zatem możemy stwierdzić, że nasz zbiór danych posiada najwięcej klientów, którzy wydają przeciętne wartości podczas rozmów telefonicznych, których czas również nie jest ani za krótki, ani za długi. Wartości centrów skupień zostały również przedstawione na wykresie (Rysunek 8).



Rysunek 8 – Wykres przedstawiający centra skupień dla zgrupowanych metodą k-Means klientów ze zbioru danych klienci3

Wynik grupowania dla przedstawicieli ze zbioru danych klienci3 został zaprezentowany na Rysunku 9.

Cluster Model

```
Cluster 0: 1 items
Cluster 1: 1 items
Cluster 2: 1 items
Cluster 3: 1 items
Cluster 4: 1 items
Total number of items: 5
```

Rysunek 9 - Wynik grupowania dla przedstawicieli ze zbioru danych klienci3 (dla k=5)

Jak widać powyżej klienci zostali podzieleni na 5 grup i w każdej z nich znalazł się jeden przedstawiciel. Taki wynik nie jest dla nas w żaden sposób użyteczny dlatego należało zmienić liczbę grup jaka ma być utworzona np. na 3 i wówczas mogliśmy zobaczyć jak przydzieleni zostali poszczególni przedstawiciele.

Cluster Model

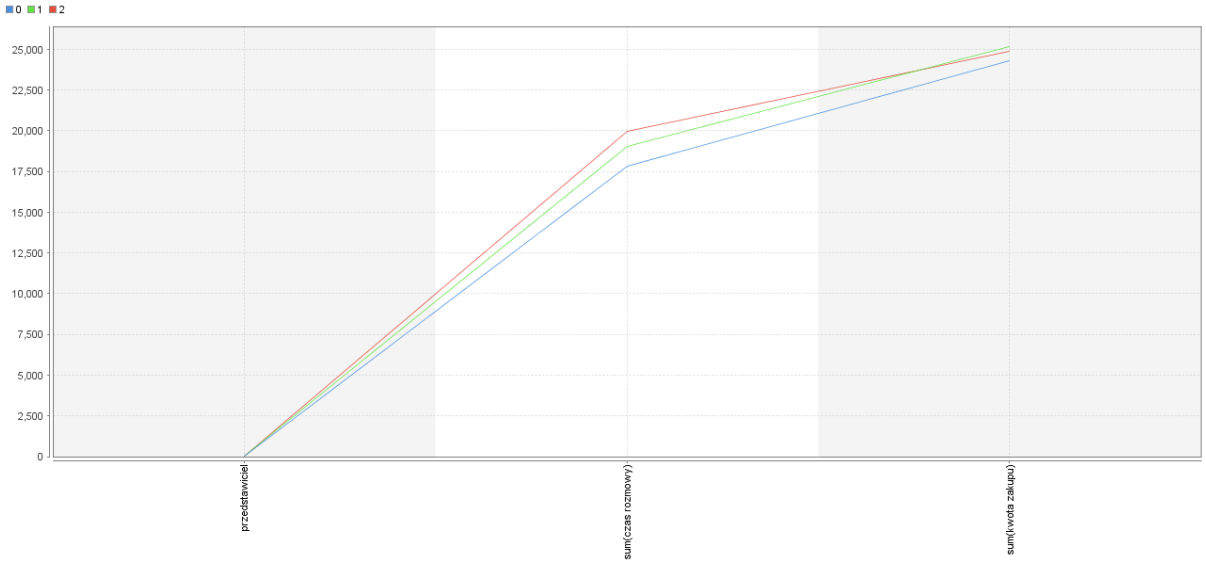
```
Cluster 0: 3 items
Cluster 1: 1 items
Cluster 2: 1 items
Total number of items: 5
```

Rysunek 10 - Wynik grupowania dla przedstawicieli ze zbioru danych klienci3 (dla k=3)

Widzimy, że przy podziale na 3 grupy najliczniejszy zbiór stanowi Cluster 0 – 3 przedstawicieli. Pozostałe grupy zwierają po jednym przedstawicielu. Centra skupień wówczas kształtują się następująco:

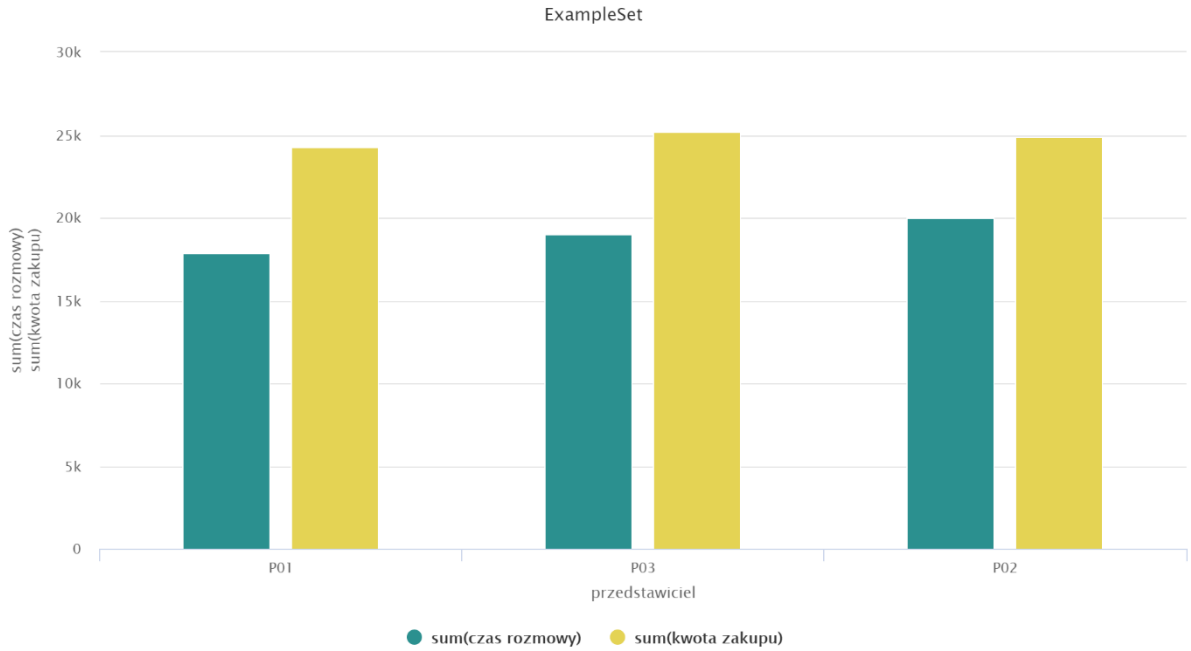
Attribute	cluster_0	cluster_1	cluster_2
przedstawiciel	4	1	2
sum(czas rozmowy)	17857.333	19022	19976
sum(kwota zakupu)	24290.667	25155	24857

Tabela 6 - Centra skupień dla zgrupowanych metodą k-Means przedstawicieli ze zbioru danych klienci3



Rysunek 11 - Centra skupień dla zgrupowanych metodą k-Means przedstawicieli ze zbioru danych klienci3

Jak widać powyżej przedstawiciele zostali zgrupowani nieco inaczej niż klienci, ponieważ Cluster 2 stanowią przedstawiciele o najwyższym sumarycznym czasie rozmowy jednak nie o najwyższej kwocie, ponieważ ta przypadła dla Cluster 1. Cluster 0 grupuje natomiast przedstawicieli o najniższej sumarycznej kwocie zakupu oraz czasie rozmowy, co możemy także dostrzec na wykresie przedstawionym na Rysunku 12.



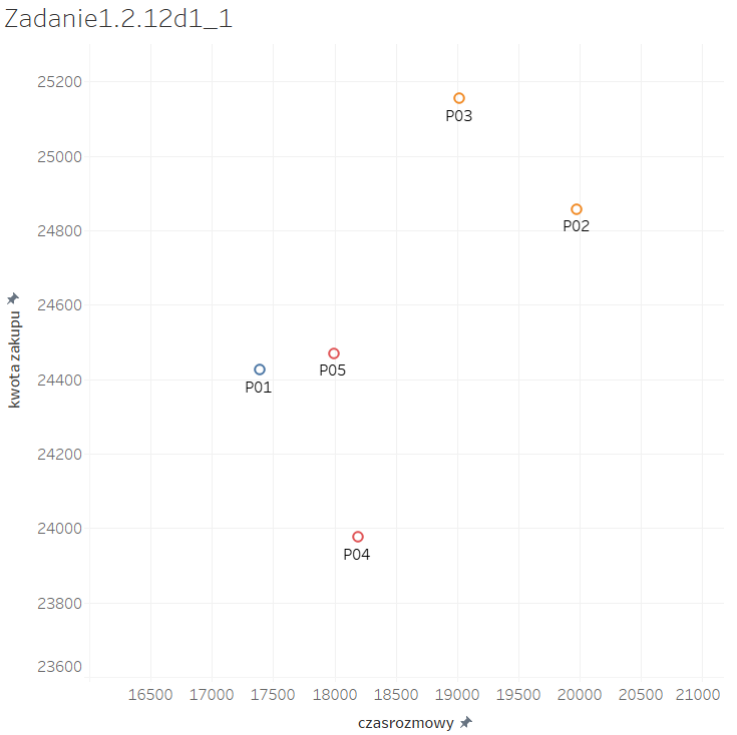
Rysunek 12 - Wykres przedstawiający centra skupień dla zgrupowanych metodą k-Means przedstawicieli ze zbioru danych klienci3

Porównując wyniki grupowania otrzymane za pomocą RapidMiner'a oraz Tableau możemy dostrzec pewne różnice.

Rysunek 13 - Wyniki klasyfikacji klientów w Tableau

Jak widzimy powyżej w Tableau otrzymaliśmy 5 grup jednak o zupełnie innych wielkościach, a zatem o zupełnie innych parametrach centrów skupień.

Taka sama sytuacja jest w przypadku przedstawicieli. Widzimy, że dla 3 grup otrzymaliśmy podział 1-2-2, a nie tak jak w RapidMiner 3-1-1.



Rysunek 14 - Wyniki klasyfikacji przedstawicieli w Tableau

Porównując zaś same środowiska można stwierdzić, że RapidMiner pozwala nam na prostszą i bardziej dogłębną analizę danych. Posiada on wiele funkcji, które w Tableau trzeba tworzyć samemu (jak np. klasyfikacja za pomocą algorytmu k-Means), a także umożliwia nam przedstawienie wyników na różnego rodzaju wizualizacjach.

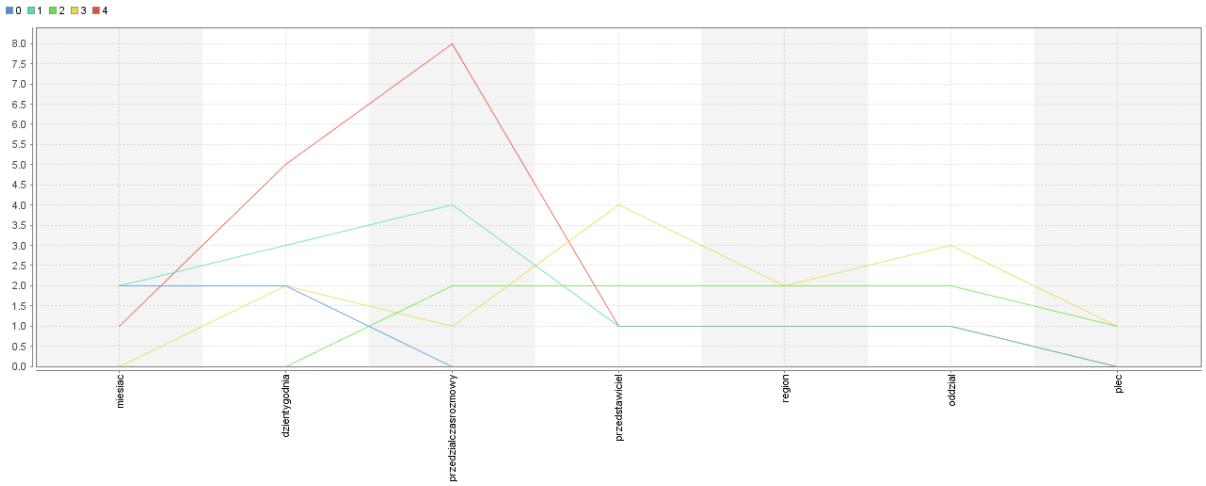
Do wszystkich poprzednich procesów zostały dodane bloki umożliwiające przeprowadzenie klasyfikacji za pomocą algorytmu k-Medoids. Proces został zapisany pod nazwą proces3.3_5. Poniżej przedstawiono wyniki poszczególnych klasyfikacji.

Cluster Model

```
Cluster 0: 251 items
Cluster 1: 180 items
Cluster 2: 228 items
Cluster 3: 263 items
Cluster 4: 78 items
Total number of items: 1000
```

Rysunek 15 - Wyniki grupowania danych o klientach metodą k-Medoids

Jak możemy zauważyć na Rysunku 15 wyniki grupowania metodą k-Medoids różnią się od tych otrzymanych dzięki algorytmowi k-Means, a jak możemy zobaczyć na poniższym wykresie najbardziej zróżnicowane wartości dla centrów skupień wystąpiły dla pola przedzialkwotowy czy dzientygodnia.



Rysunek 16 - Wykres przedstawiający wartości centrów skupień dla zgrupowanych danych o klientach metoda k-Medoids

Rozbieżności wystąpiły również dla danych o klientach bez użycia numeru klienta oraz dla zagregowanych wartości kwoty zakupu oraz czasów rozmów dla klientów, natomiast wynik dla przedstawicieli jest bardzo zbliżony do tego otrzymanego przy użyciu k-Means.

Cluster Model

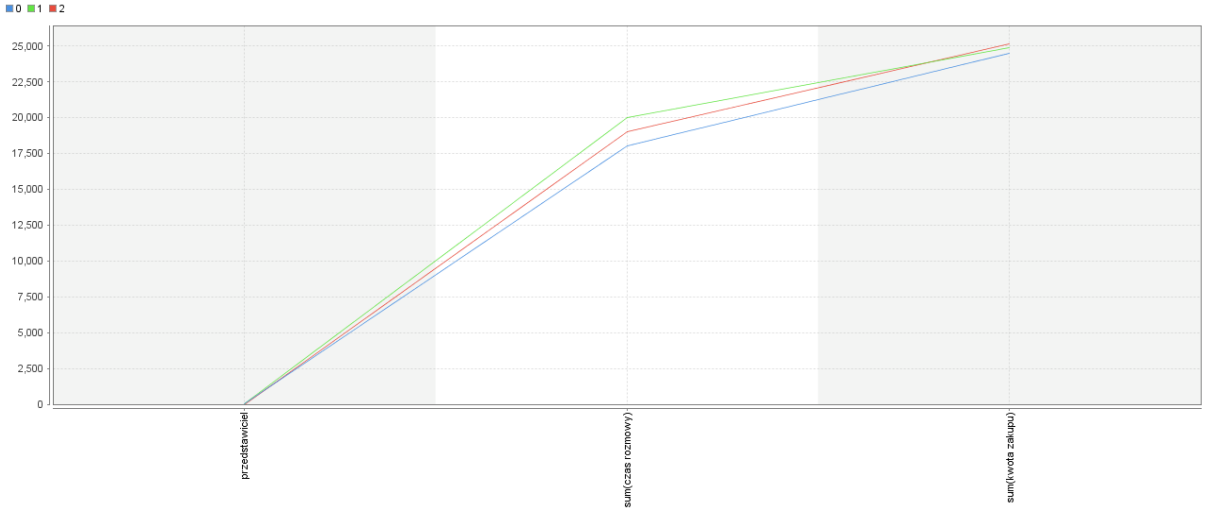
```
Cluster 0: 3 items
Cluster 1: 1 items
Cluster 2: 1 items
Total number of items: 5
```

Rysunek 17 - Wynik grupowania dla przedstawicieli ze zbioru danych klienci3 przy użyciu algorytmu k-Medoids

Centra skupień prezentują się wówczas w następujący sposób:

Attribute	cluster_0	cluster_1	cluster_2
przedstawiciel	2	1	0
sum(czas rozmowy)	17992	19976	19022
sum(kwota zakupu)	24470	24857	25155

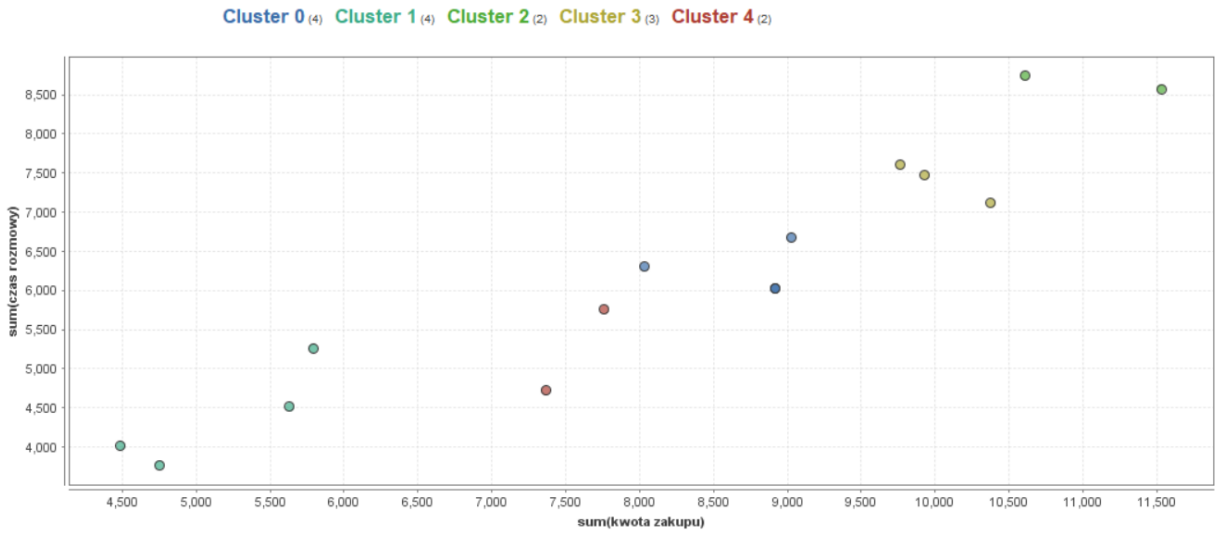
Tabela 7 - Centra skupień dla zgrupowanych metodą k-Medoids przedstawicieli ze zbioru danych klienci3



Rysunek 18 - Centra skupień dla zgrupowanych metodą k-Medoids przedstawicieli ze zbioru danych klienci3

Jak widać wartości dla centrów skupień w przypadku przedstawicieli są bardzo zbliżone do tych otrzymanych algorytmem k-Means, natomiast wizualizacje są niemalże identyczne.

Wszystkie otrzymane wyniki zostały zapisane do pliku wyniki_proces3.3.res za pomocą bloku Write as Text. Dodatkowo wyniki klasyfikacji dla poszczególnych danych i zastosowanych algorytmów zostały wyeksportowane do plików .xlsx. Dodany został także blok ClusterModelVisualizer dla zgrupowanych klientów oraz przedstawicieli metodą k-Means. Pozwoliło nam to uzyskać m.in. wizualizację podobną do tych otrzymanych w Tableau.



Rysunek 19 - Wynik klasyfikacji klientów w RapidMiner

Dzięki takiej wizualizacji możemy w szybki sposób odczytać do jakiej grupy dany klient bądź przedstawiciel został zakwalifikowany.

Rysunek 20 - Wynik klasyfikacji przedstawicieli w RapidMiner

3.3.2 optymalizacja parametrów modelu klasyfikacyjnego

Korzystając z wbudowanych, przykładowych modeli w RapidMiner, został wczytany ten o nazwie Credit Risk Modelling. Proces ten zawiera takie bloki jak Log to Data, dzięki któremu możemy zobaczyć wartości obliczone podczas procesu oraz Optimize Parameters, który zwraca optymalny zestaw parametrów w celu uzyskania jak najlepszych wyników. Wyniki procesu prezentuje Rysunek 21.

iteration	SVM.kernel_gamma	SVM.C	accuracy ↓
46	0.019	10.000	0.956
47	0.035	10.000	0.954
57	0.019	100	0.954
45	0.010	10.000	0.951
56	0.010	100	0.951
36	0.035	1.000	0.951
37	0.065	1.000	0.951
58	0.035	100	0.949
59	0.065	100	0.949
34	0.010	1.000	0.946
35	0.019	1.000	0.946
48	0.065	10.000	0.946
60	0.120	100	0.923
49	0.120	10.000	0.921

Rysunek 21 – Fragment wyników optymalizacji parametrów dla modelu Credit Risk Modelling

Powyżej widzimy, że dla wartości parametrów kernel_gamma równej 0,019 oraz C równego 10 uzyskujemy najlepszą poprawność klasyfikacji na poziomie 0,956.

Do parametrów został dodany karnel_type o typach radial i annova. Wyniki po zmianie zostały zaprezentowane na Rysunku 22.

iteration	SVM.kernel_gamma	SVM.C	SVM.kernel_type	acc... ↓
84	0.416	0.010	anova	0.959
81	0.065	0.010	anova	0.956
89	0.010	0.100	anova	0.956
46	0.019	10.000	radial	0.956
87	2.686	0.010	anova	0.956
121	5	10.000	anova	0.956
132	5	100	anova	0.956
47	0.035	10.000	radial	0.954
108	1.443	1.000	anova	0.954
110	5	1.000	anova	0.954
57	0.019	100	radial	0.954
74	0.775	0.001	anova	0.954
93	0.120	0.100	anova	0.954
131	2.686	100	anova	0.954
45	0.010	10.000	radial	0.951
56	0.010	100	radial	0.951
75	1.443	0.001	anova	0.951

Rysunek 22 - Fragment wyników optymalizacji parametrów dla Credit Risk Modelling po dodaniu parametru karnel_type

Jak widzimy wyniki poprawiły się w bardzo niewielkim stopniu, poprawność klasyfikacji wzrosła z 0,956 na 0,959. Taki wynik osiągniemy, jeżeli jako kernel_type użyjemy annova.

W bloku Optimize Parameters zastąpiono poprzedni klasyfikator algorytmem Decision Tree. Wyniki procesu zostały przedstawione na Rysunku 23.

iteration	Decision Tree.minimal_leaf_size	Decision Tree.criterion	acc... ↓
23	1	accuracy	0.946
4	31	gain_ratio	0.941
5	41	gain_ratio	0.938
2	11	gain_ratio	0.938
3	21	gain_ratio	0.938
6	51	gain_ratio	0.938
7	60	gain_ratio	0.938
17	51	information_gain	0.938
25	21	accuracy	0.938
26	31	accuracy	0.938

Rysunek 23 – Fragment wyników optymalizacji parametrów dla Credit Risk Modelling dla klasyfikatora Decision Tree

Porównując wyniki do tych otrzymanych poprzednio widzimy, że poprawność klasyfikacji jest mniejsza, jednak jest to niewielka różnica.

Sprawdzając czy optymalizacja parametrów klasyfikatora DecisionTree dla danych klienci6.arff i klienci6new.arff może poprawić jakość wyników użyto bloków analogicznych do tych zastosowanych w przykładzie Credit Risk Modeling. Wyniki przeprowadzonego procesu przedstawiono poniżej:

iteration	Decision Tree.minimal_leaf_size	Decision Tree.criterion	acc... ↓
13	11	information_gain	0.480
25	21	accuracy	0.479
1	1	gain_ratio	0.476

Rysunek 24 - Fragment wyników optymalizacji parametrów dla danych o klientach z zadania 2.4

Jak widać, najlepszą jakość klasyfikacji uzyskamy dla criterion information_gain oraz dla minimal_leaf_size równego 11. Poprawność klasyfikacji będzie wynosiła wówczas 0,48 zatem nie możemy mówić tutaj o wysokiej skuteczności klasyfikacji. Poprzedni wynik uzyskany dla domyślnych wartości parametrów wynosił 0,476 zatem widzimy, że istnieje możliwość jego poprawy.