Projekt 2- Klasteryzacja

Krzysztof Sawicki, Natalia Safiejko, Michał Geneja

Preprocessing

Zapoznanie się ze zbiorem danych

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown

Buisness case: podział klientów na grupy społeczne co może pomóc w wybraniu odpowiedniej strategii marketingu

Transformacja data frame

1. Zamiana słów na liczby
2. Pozbycie się outlierów
3. Normalizacja
4. Usunięcie skorelowanych kolumn
5. Zmiana wag zawodów

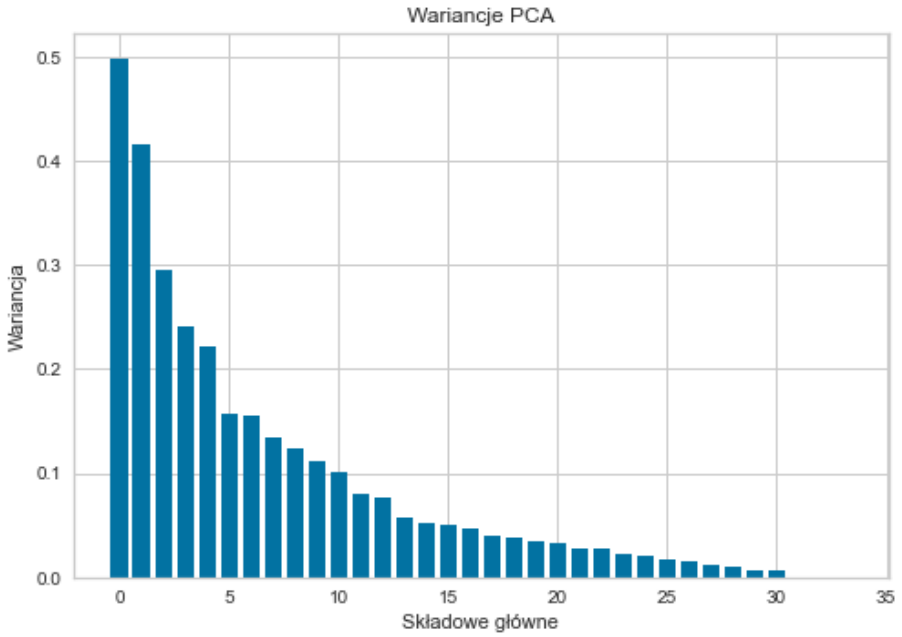
	age	default	balance	housing	loan	day	month	duration
0	0.467532	0.0	0.091908	0.0	0.0	0.833333	0.636364	0.221274
1	0.272727	0.0	0.044147	0.0	0.0	0.866667	0.636364	0.091267
2	0.233766	0.0	0.059292	0.0	0.0	0.433333	0.636364	0.067630
3	0.233766	0.0	0.722098	0.0	0.0	0.266667	0.454545	0.084045
4	0.246753	0.0	0.057214	1.0	1.0	0.500000	0.454545	0.139199

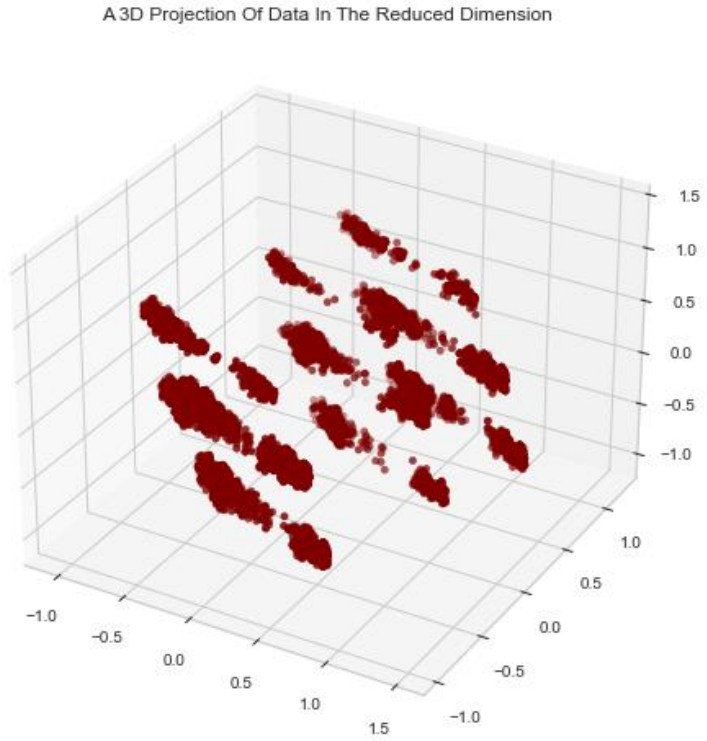
duration	campaign	pdays	...	marital_single	education_primary	education_secondary	education_tertiary
0.221274	0.15	0.0	...	0.0	0.0	1.0	0.0
0.091267	0.15	0.0	...	1.0	0.0	0.0	1.0
0.067630	0.05	0.0	...	1.0	0.0	0.0	1.0
0.084045	0.00	0.0	...	1.0	0.0	1.0	0.0
0.139199	0.00	0.0	...	0.0	0.0	0.0	1.0

education_unknown	contact_cellular	contact_telephone	poutcome_failure	poutcome_other	poutcome_success
0.0	1.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0

Clustering

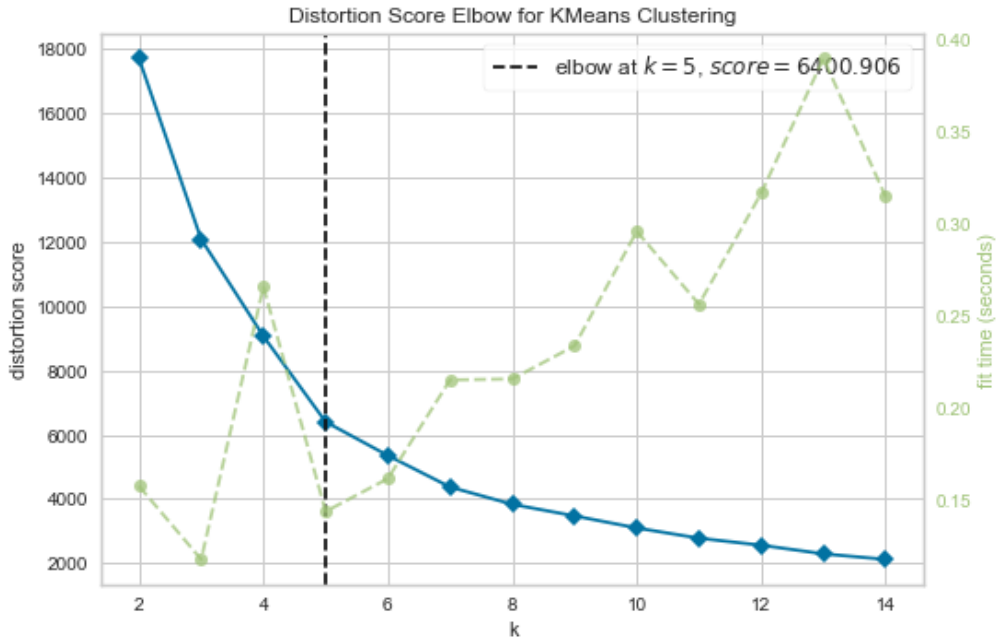
Wykorzystanie PCA



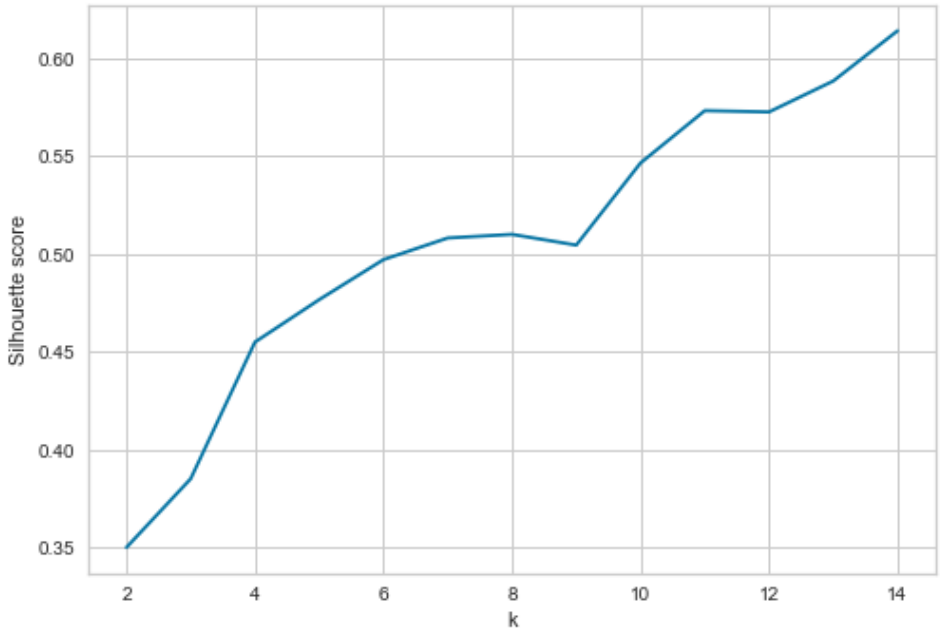


Poszukiwanie optymalnej liczby klastrów

Kmeans

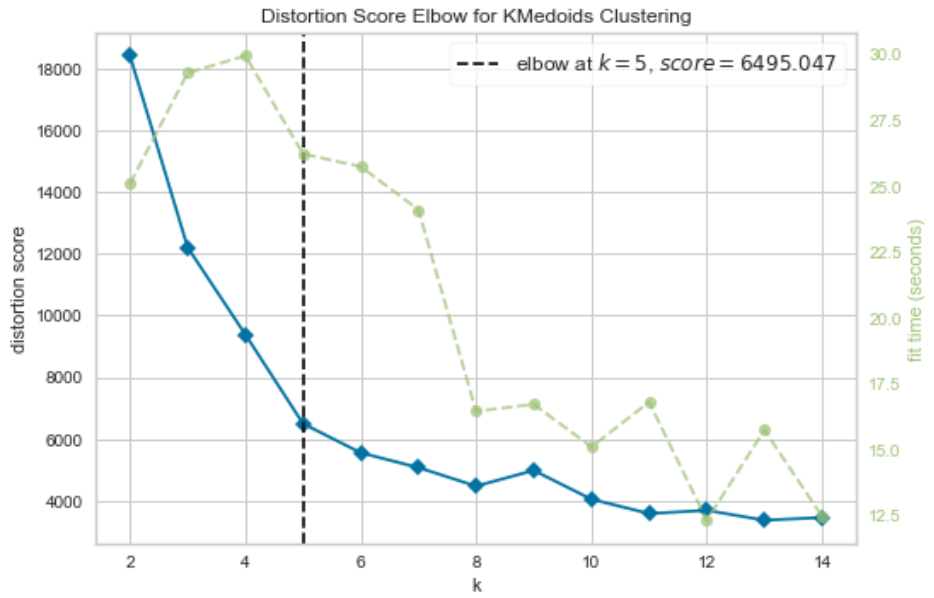


Silhouette score

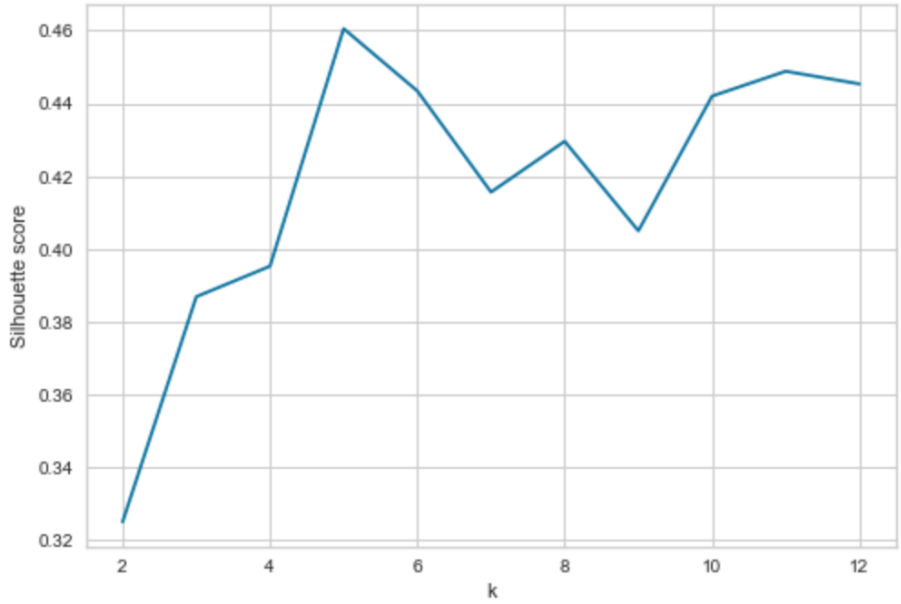


Poszukiwanie optymalnej liczby klastrów

Kmedoids

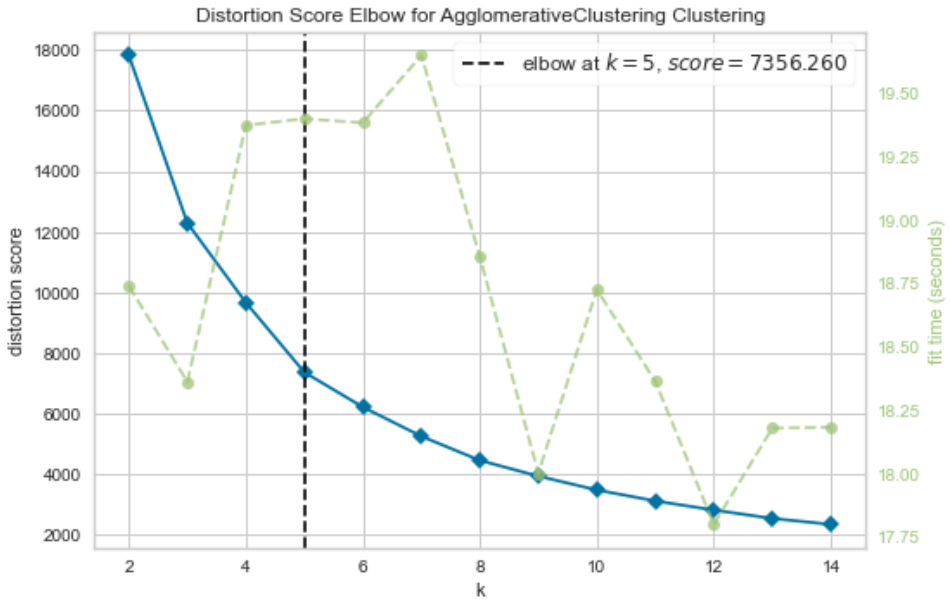


Silhouette score

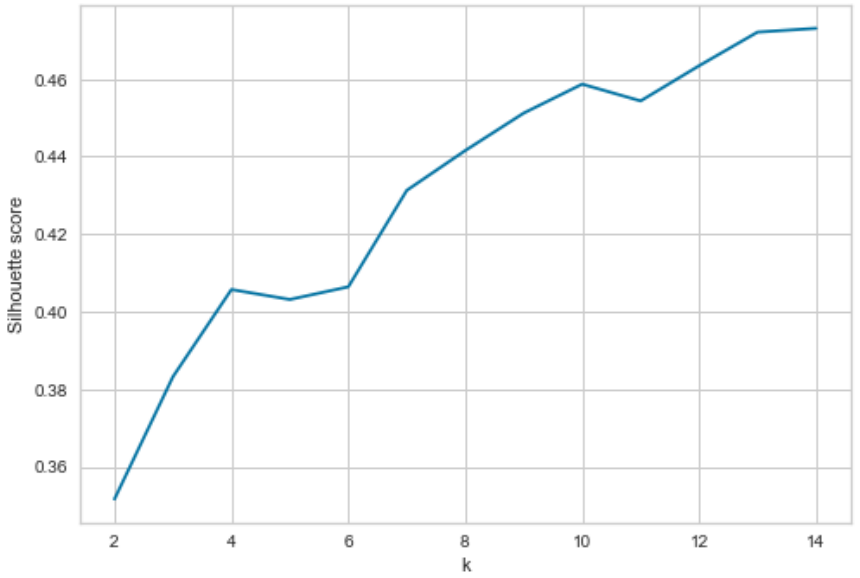


Poszukiwanie optymalnej liczby klastrów

Agglomerative Clustering

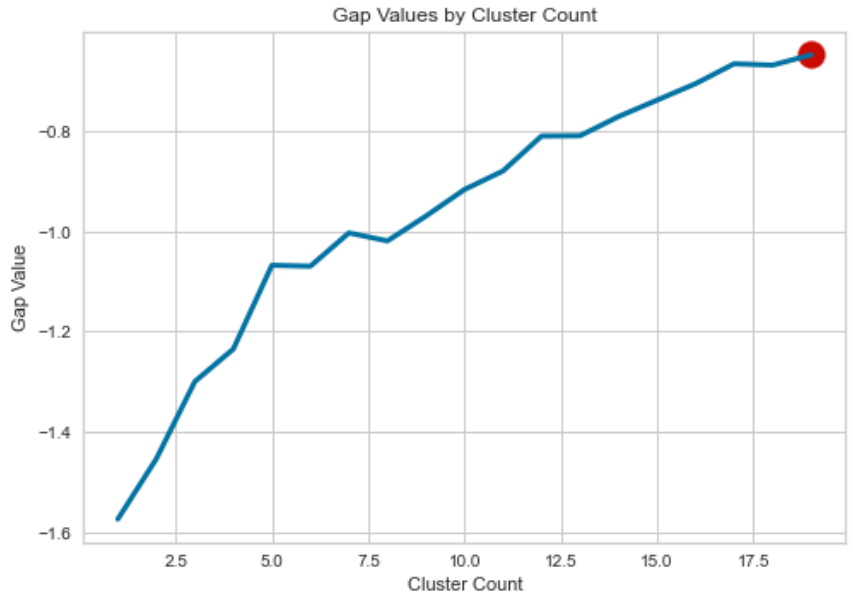


Silhouette score

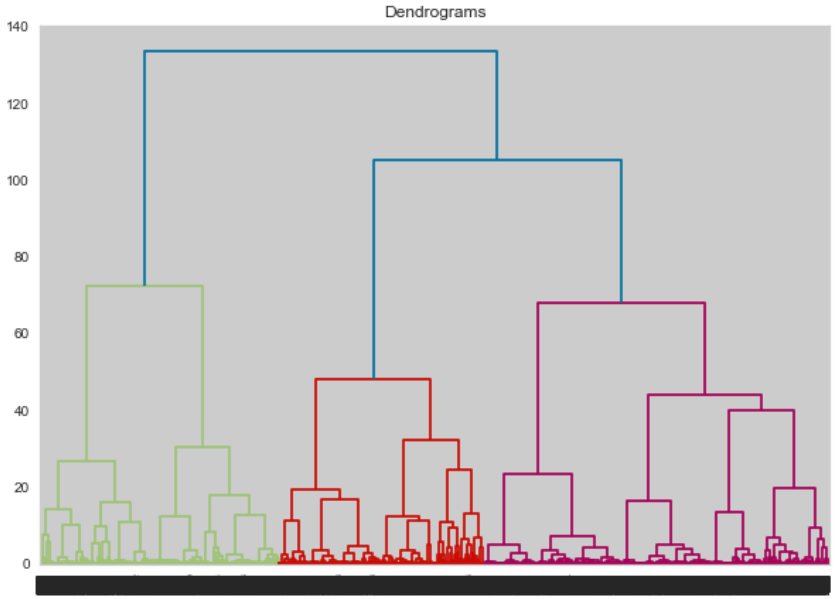


Poszukiwanie optymalnej liczby klastrów

Gap Statistics

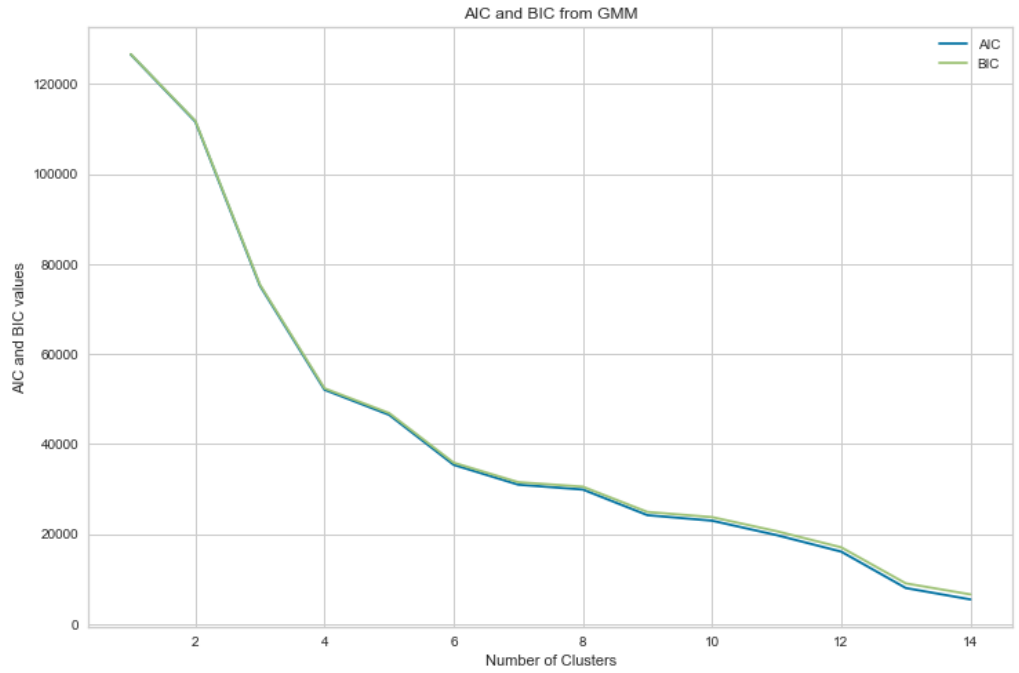


Dendogram



Poszukiwanie optymalnej liczby klastrów

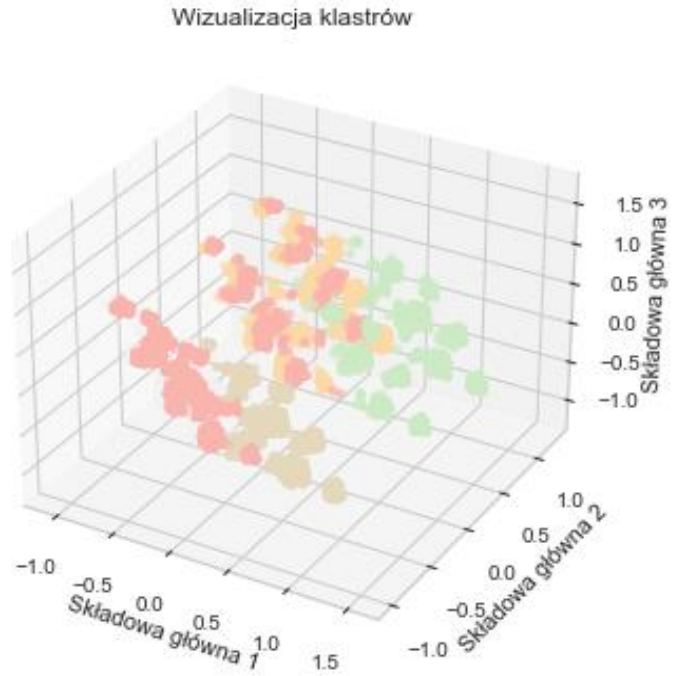
AIC and BIC from GMMs



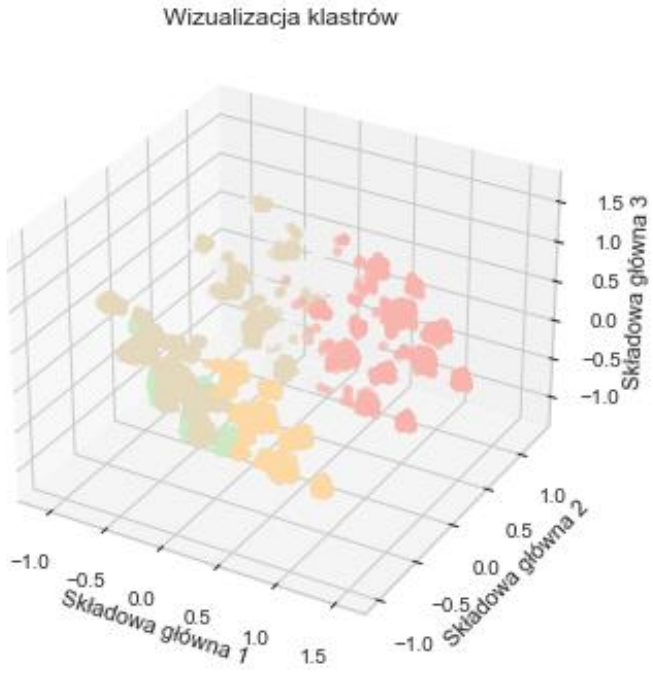
Wnioski:
Większość testów wskazywało na przedział 4-6 lub na jak największą ilość klastrów. Za optymalną ilość klastrów przyjmujemy 5

Klasteryzacja

Kmeans

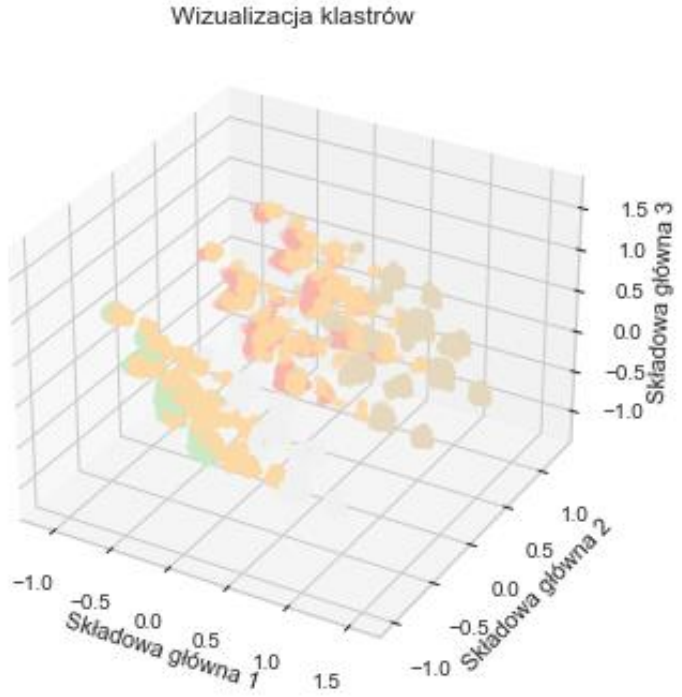


MiniBatchKMeans

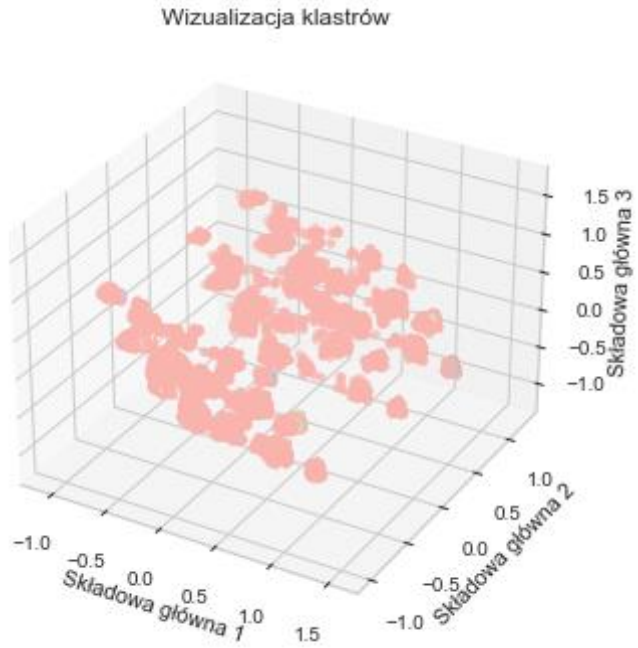


Klasteryzacja

Agglomerative Clustering



DBSCAN



Wyniki- wybór najlepszego modelu

	Metoda	Silhouette Score	Calinski-Harabasz Indedf
0	KMeans	0.161599	2683.451788
1	MiniBatchKMeans	0.133408	2166.216575
2	Agglomerative Clustering	0.142821	2387.008756
3	DBSCAN	-0.176301	187.736549

Charakterystyka każdego z klastrów- mediany

	age	default	balance	housing	loan	day	month	duration	campaign	pdays	job_admin.	job_blue-collar	job_entrepreneur
0	41.0	0.0	594.0	0.0	0.0	17.0	7.0	170.0	2.0	-1.0	0.0	0.000000	0.0
1	45.0	0.0	404.5	1.0	0.0	15.0	6.0	176.0	2.0	-1.0	0.0	1.428571	0.0
2	34.0	0.0	339.0	1.0	0.0	16.0	5.0	184.0	2.0	-1.0	0.0	0.000000	0.0
3	41.0	0.0	453.0	1.0	0.0	16.0	6.0	179.0	2.0	-1.0	0.0	0.000000	0.0
4	34.0	0.0	564.0	0.0	0.0	17.0	6.0	183.0	2.0	-1.0	0.0	0.000000	0.0

job_housemaid	job_management	job_retired	job_self-employed	job_services	job_student	job_technician
0.0	1.428571	0.0	0.0	0.0	0.0	0.0
0.0	0.000000	0.0	0.0	0.0	0.0	0.0
0.0	0.000000	0.0	0.0	0.0	0.0	0.0
0.0	0.000000	0.0	0.0	0.0	0.0	0.0
0.0	1.428571	0.0	0.0	0.0	0.0	0.0

```
Cluster 1 has 6796 points.
Cluster 3 has 3129 points.
Cluster 2 has 4886 points.
Cluster 4 has 3639 points.
Cluster 0 has 3702 points.
```

	job_unemployed	job_unknown	marital_divorced	marital_married	marital_single	education_primary	education_secondary	education_tertiary	education_unknown
0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
1	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
4	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0

Charakterystyka każdego z klastrów- średnie

	age	default	balance	housing	loan	day	month	duration	campaign	pdays	job_admin.
0	42.602913	0.013190	1605.744160	0.451772	0.152240	16.032152	6.700192	245.537236	2.747733	30.774938	0.048286
1	46.088679	0.014825	1197.667655	0.571968	0.139623	15.228841	5.950674	245.513208	2.703504	34.121563	0.049288
2	36.353584	0.026487	943.577044	0.607105	0.167332	15.517974	5.796300	256.408661	2.495060	42.434307	0.306316
3	42.412301	0.017510	1176.533549	0.609476	0.192908	15.946881	6.152884	247.241613	2.719688	38.170541	0.229126
4	36.553231	0.014462	1479.520308	0.468000	0.114769	16.126154	6.210769	259.372000	2.641231	36.212923	0.074286

job_blue-collar	job_entrepreneur	job_housemaid	job_management	job_retired	job_self-employed	job_services	job_student	job_technician
0.016095	0.087544	0.026302	0.850705	0.057708	0.088721	0.023947	0.005496	0.183724
0.768964	0.041586	0.127840	0.058529	0.151328	0.025799	0.073161	0.009241	0.046592
0.271480	0.024926	0.017418	0.059762	0.030031	0.021022	0.219526	0.077179	0.354065
0.375011	0.040570	0.027537	0.071891	0.082401	0.042672	0.203691	0.001892	0.306483
0.020659	0.049231	0.013187	0.786813	0.021978	0.076923	0.029451	0.072527	0.236923

job_unemployed	job_unknown	marital_divorced	marital_married	marital_single	education_primary	education_secondary	education_tertiary	education_unknown
0.027873	0.012170	0.000000	1.000000	0.000000	0.000000	0.000000	0.949986	0.050014
0.052753	0.023489	0.112938	0.775741	0.111321	0.907008	0.000000	0.000000	0.092992
0.042043	0.004805	0.290940	0.000000	0.709060	0.000000	0.972462	0.000000	0.027538
0.041831	0.005465	0.000000	1.000000	0.000000	0.000000	0.993673	0.000000	0.006327
0.036923	0.009670	0.231692	0.000000	0.768308	0.000308	0.000000	0.946769	0.052923

Charakterystyka każdego z klastrów - zbiór testowy

	age	default	balance	housing	loan	day	month	duration	campaign	pdays	job_admin.	job_blue-collar	job_entrepreneur
0	45.852954	0.024945	1169.721663	0.559737	0.139606	15.445077	6.016193	250.301969	2.704158	31.315974	0.035011	0.550985	0.029759
1	36.477284	0.017356	1403.685043	0.454313	0.108729	15.955079	6.207759	258.386932	2.537519	36.377744	0.054109	0.017356	0.041858
2	42.837511	0.017010	1554.205013	0.469561	0.139660	16.244405	6.603850	238.220233	2.742614	34.757386	0.041629	0.004029	0.061325
3	42.306393	0.013752	1185.858142	0.622678	0.199517	15.910736	6.144029	249.268758	2.645356	38.960193	0.155368	0.262485	0.033293
4	36.974498	0.020401	1004.445767	0.586875	0.171710	15.752465	5.865692	257.246175	2.440666	40.279157	0.217273	0.176471	0.009521

job_housemaid	job_management	job_retired	job_self-employed	job_services	job_student	job_technician
0.081838	0.039387	0.115098	0.017943	0.045514	0.006565	0.028884
0.008678	0.520674	0.019398	0.072996	0.017866	0.052067	0.160796
0.015667	0.603850	0.047001	0.061773	0.015667	0.004029	0.121307
0.020989	0.048010	0.056212	0.029916	0.145476	0.001689	0.213752
0.013941	0.045903	0.025502	0.019721	0.157429	0.046923	0.248895

job_unemployed	job_unknown	marital_divorced	marital_married	marital_single	education_primary	education_secondary	education_tertiary	education_unknown
0.035449	0.013567	0.099781	0.787309	0.112910	0.894967	0.000000	0.000000	0.105033
0.029607	0.004594	0.224094	0.000000	0.775906	0.000510	0.000000	0.935171	0.064319
0.015667	0.008057	0.000000	1.000000	0.000000	0.000000	0.000000	0.944047	0.055953
0.030157	0.002654	0.000000	1.000000	0.000000	0.000000	0.993727	0.000000	0.006273
0.036722	0.001700	0.316559	0.000000	0.683441	0.000000	0.965998	0.000000	0.034002

Wyniki dla zbioru testowego prezentują się podobnie:

```
Cluster 4 has 2941 points.
Cluster 1 has 1959 points.
Cluster 3 has 4145 points.
Cluster 0 has 2285 points.
Cluster 2 has 2234 points.
```

Podsumowanie klastrów

- Klaster 0: wykształcenie wyższe, wyższe saldo, zawód: zarządzanie
- Klaster 1: pracownicy fizyczni, żonaci, kredyt mieszkaniowy, wykształcenie podstawowe
- Klaster 2: serwisanci, technicy, rozwodnicy, single z wykształceniem średnim
- Klaster 3: serwisanci, technicy, żonaci z wykształceniem średnim
- Klaster 4: najmłodsi, wykształcenie wyższe, zawód: zarządzanie, single

Zastosowanie się do uwag walidacji

- Zespół walidacyjny zwrócił nam uwagę na to, że początkowo dokonaliśmy wyboru modelu na podstawie wizualizacji przy pomocy pca, pomijając Silhouette score czy indeks Calinskiego-Harabasza.
- Po uwzględnieniu tej uwagi doszliśmy do wniosku, że sumarycznie najlepiej sprawował się algorytm KMeans co z początku nie było oczywiste.

Dziękujemy za uwagę :)

Walidacja

Klasteryzacja chorób serca

Brak walidacji krzyżowej

Pominięte zostały metryki oceny klastrów, po dodaniu okazało się, że zarówno Silhouette jak i Calinski - Harabasz są dość niskie, a ponieważ jest tylko jeden model nie bardzo można z czymś porównać.

Na wyraźny plus jest interpretacja, choć brak wiedzy (zarówno zespołu modelującego, jak i walidacyjnego) nie pozwala na skorelowanie wszystkich klastrów z konkretnymi stanami zdrowia pacjentów.