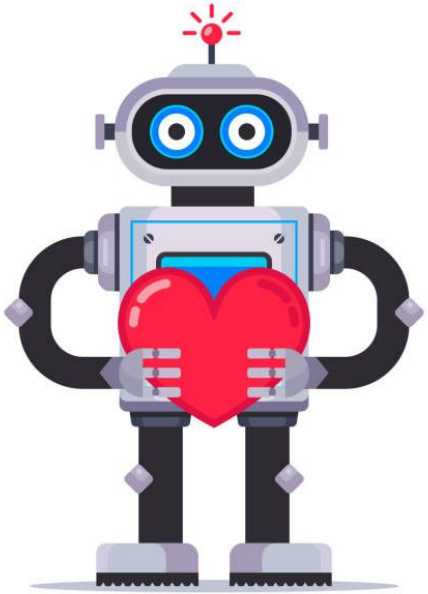
Klasteryzacja pacjentów z chorobami serca

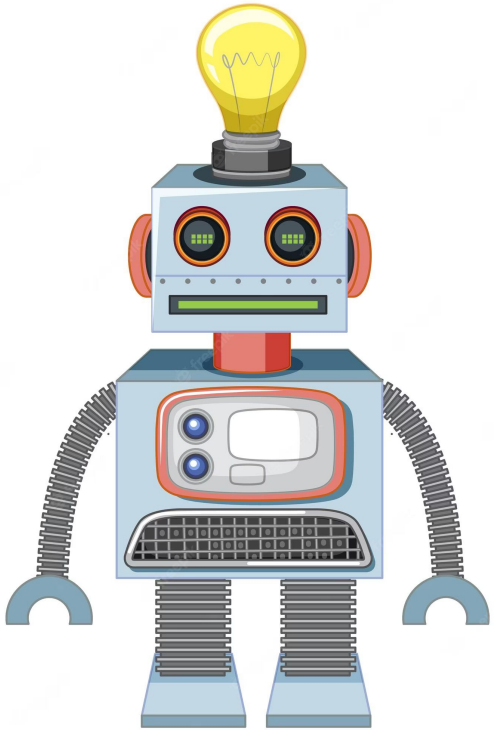
Wstęp do uczenia maszynowego

Rok akademicki 2022/2023

Wykonali: Alicja Charuza, Mateusz Gałęziewski

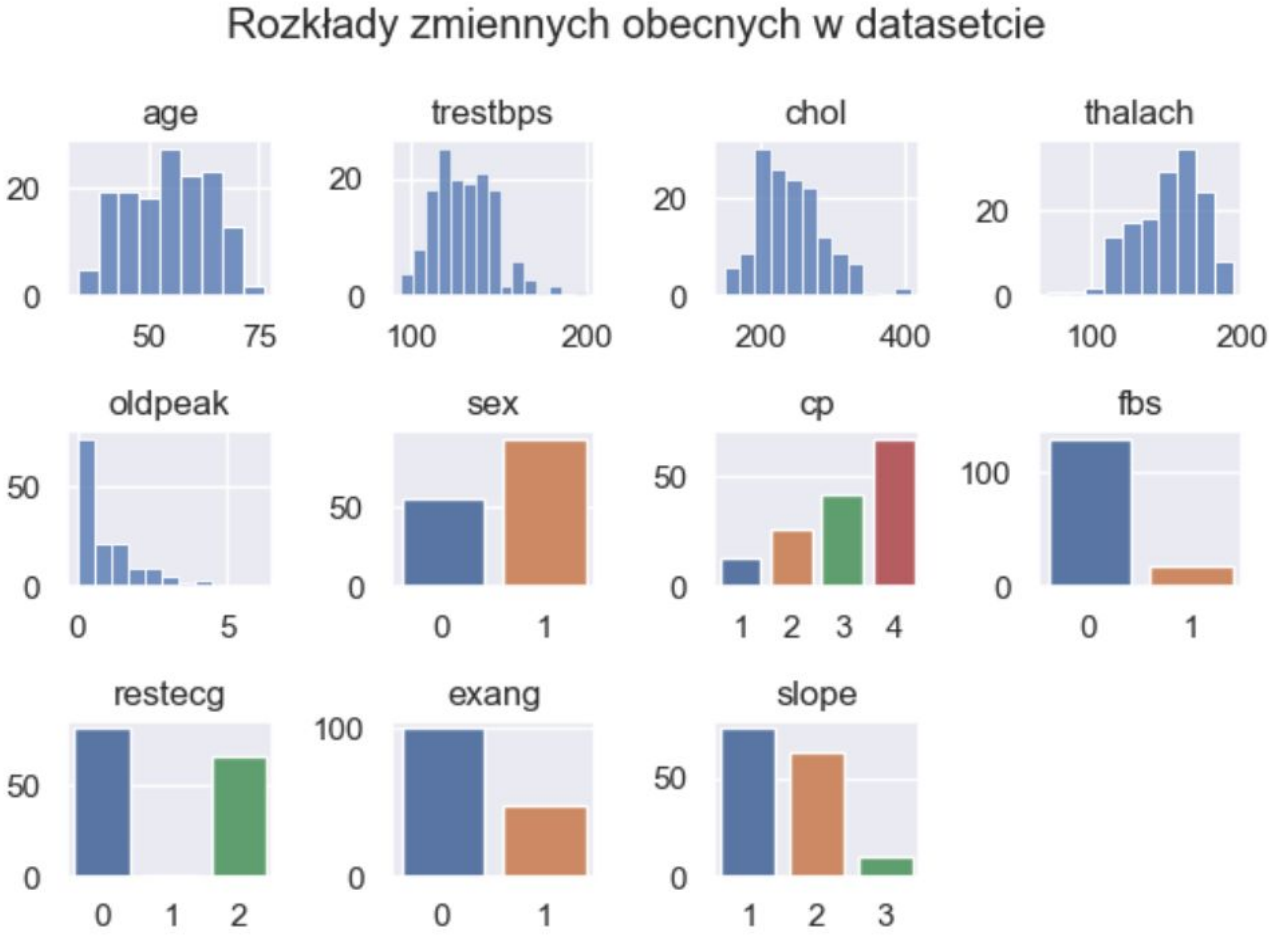


Po co grupować pacjentów ?

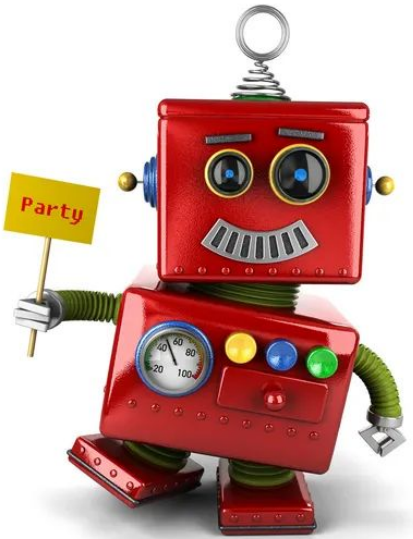


Zbiór danych i zmienne

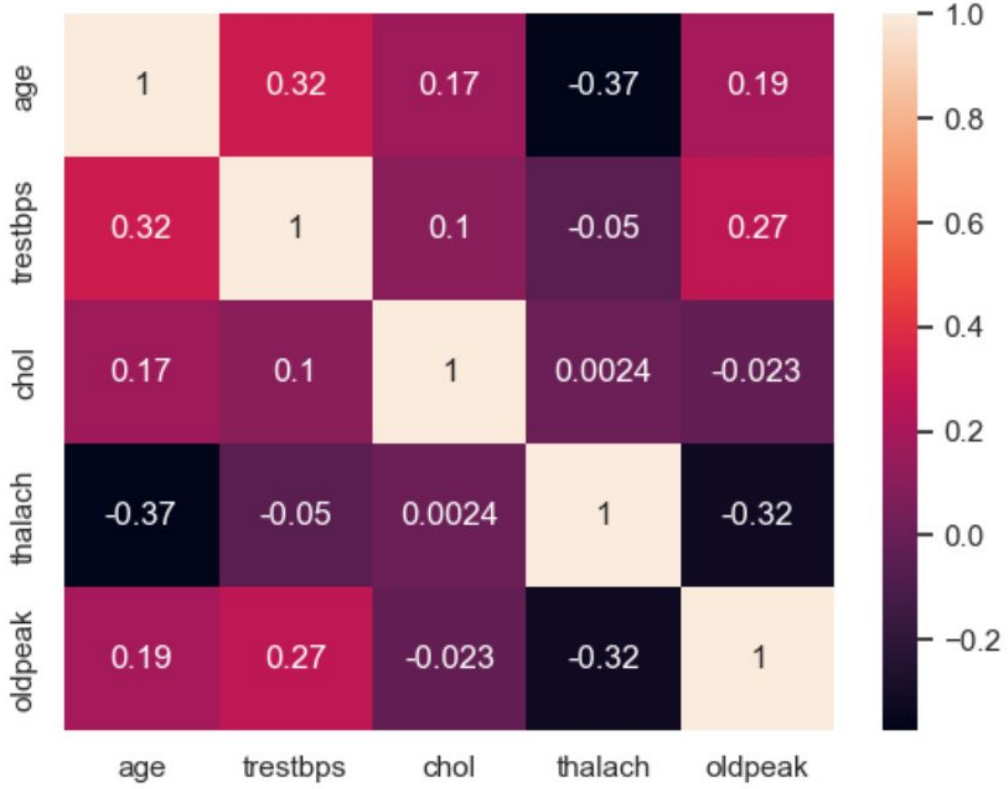
kolumna	opis	typ
id	indeks	numeryczny
age	wiek	numeryczny
sex	płeć	kategoryczny
cp	typ bólu klatki piersiowej	kategoryczny
trestbps	ciśnienie krwi w spoczynku (w momencie przyjęcia do szpitala) w mm/Hg	numeryczny
chol	poziom cholesterolu w surowicy w mg/dl	numeryczny
fbs	poziom cukru na czczo w mg/dl > 120 (1 = true, 0 = false)	kategoryczny
restecg	wyniki ekg w spoczynku	kategoryczny
thalach	maksymalne tętno	numeryczny
exang	dławica wysiłkowa (1 = true, 0 = false)	kategoryczny
oldpeak	obniżenie odcinka ST wywołane wysiłkiem w porównaniu do spoczynku	numeryczny
slope	spadek odcinka ST podczas szczytowego wysiłku	kategoryczny


Rozkłady zmiennych obecnych w datasetcie
age
trestbps
chol
thalach
oldpeak
sex
cp
fbs
restecg
exang
slope

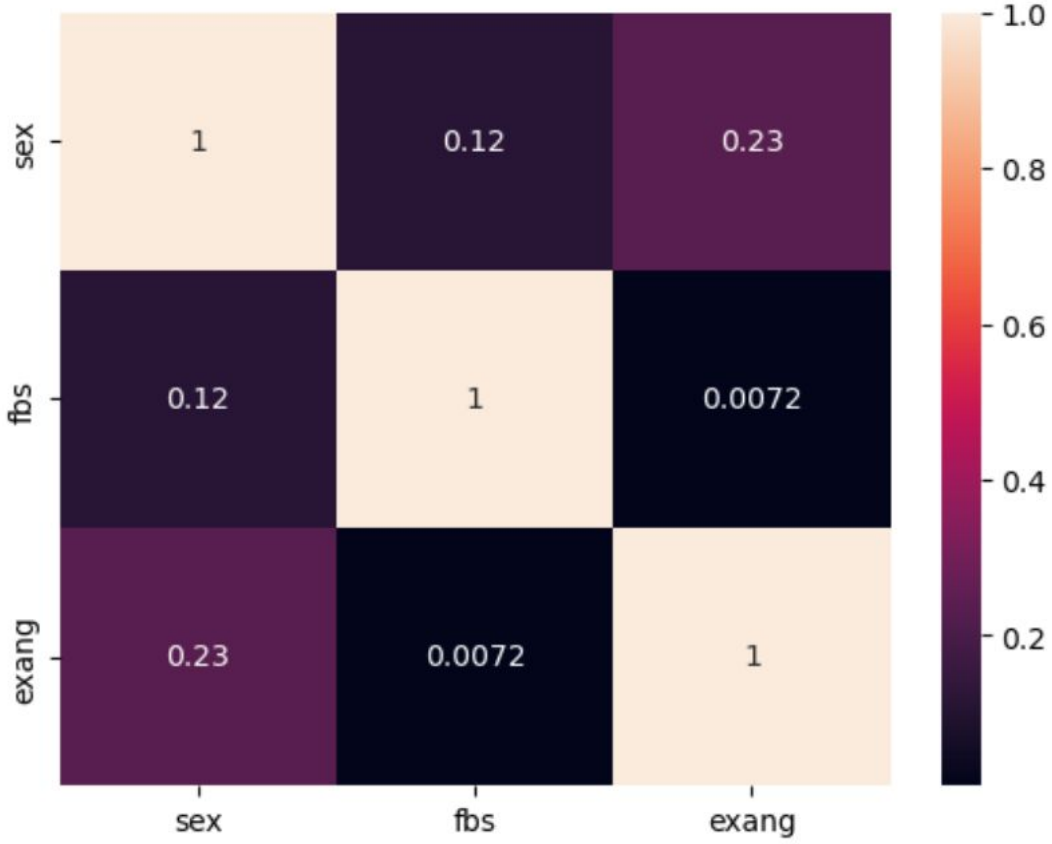
Preprocessing danych



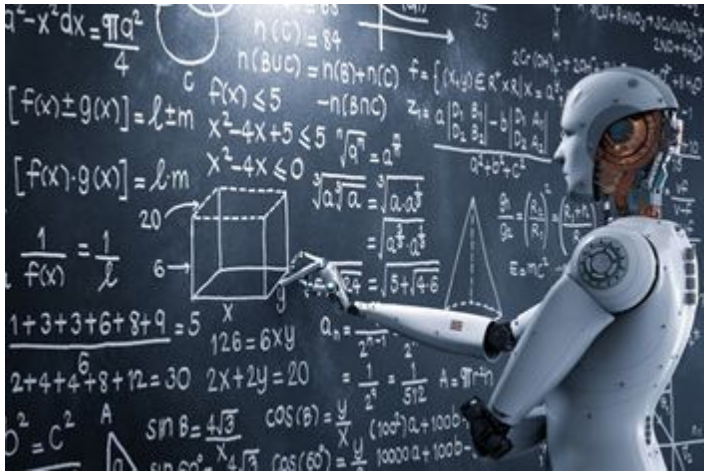
Macierz korelacji pearsona dla zmiennych ciągłych



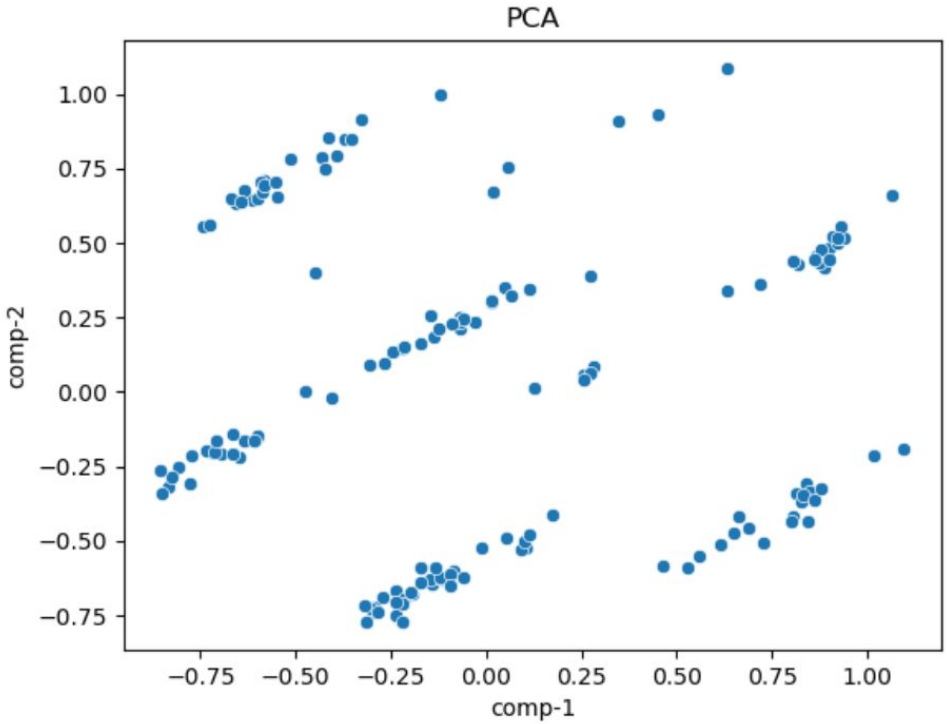
Macierz korelacji spearmana dla zmiennych dyskretnych

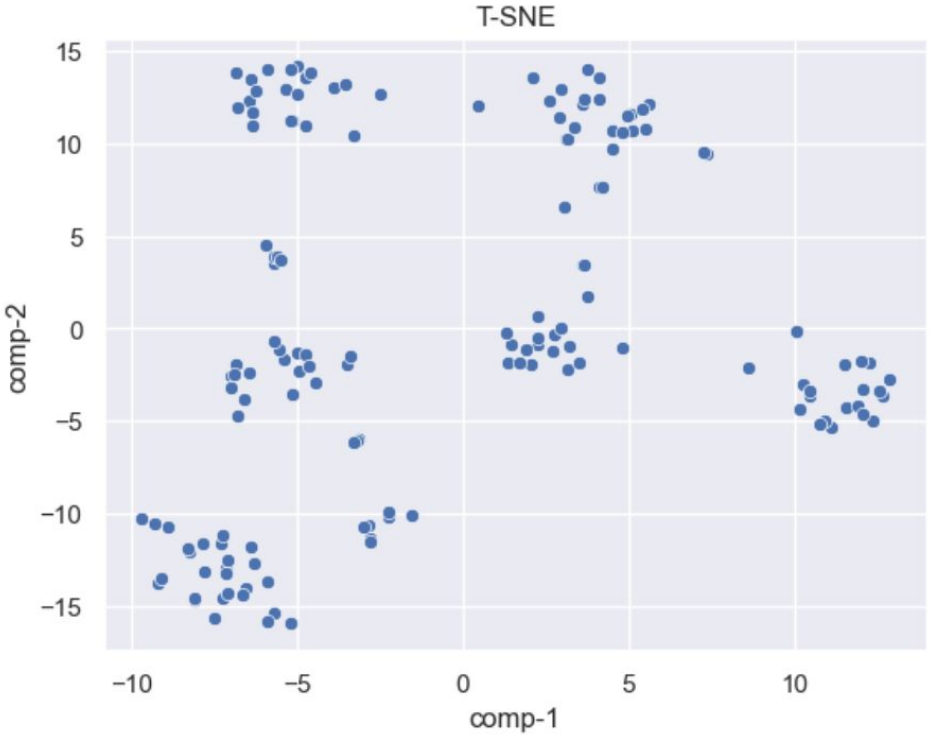


Dane zostały poddane normalizacji zawierają się w przedziale od 0 do 1

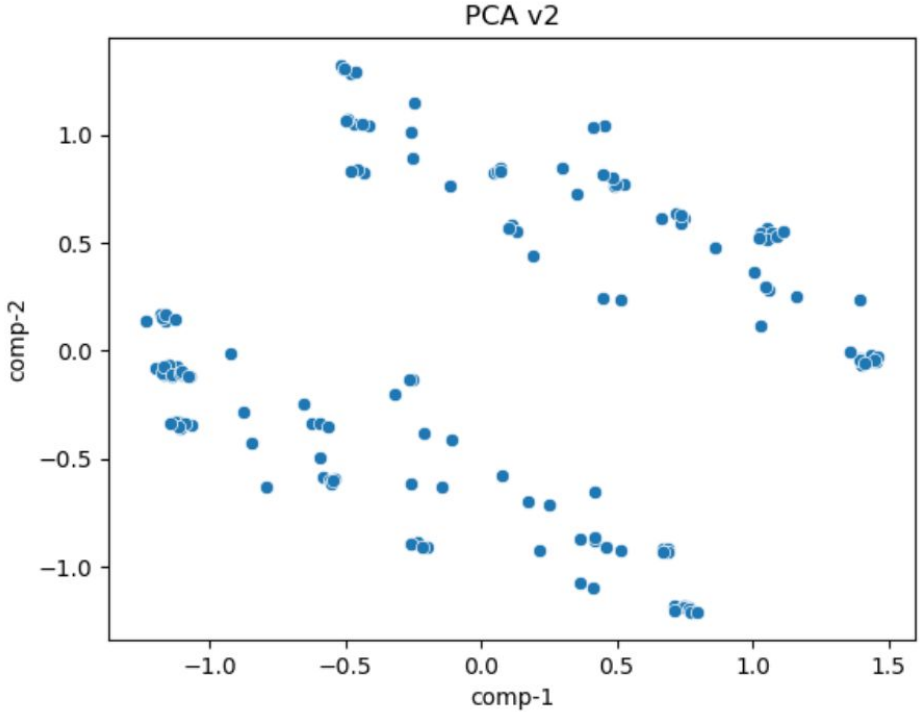


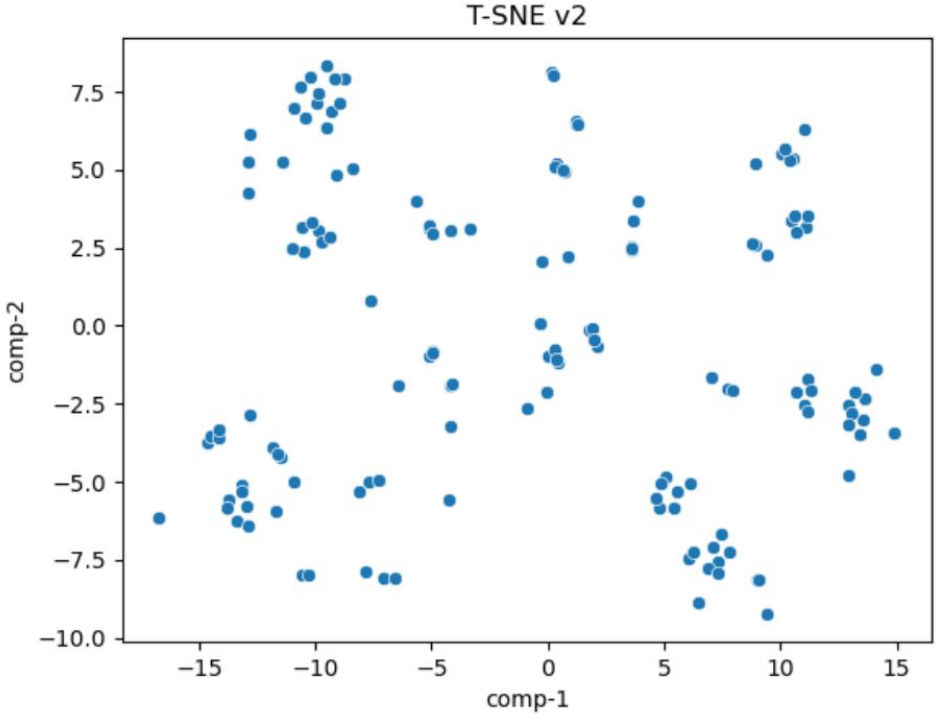
Tak wyglądają nasze dane treningowe po redukcji wymiarowości do 2 przy zastosowaniu technik PCA i t-SNE





Dlaczego zdecydowaliśmy się nie kodować zmiennych dyskretnych?





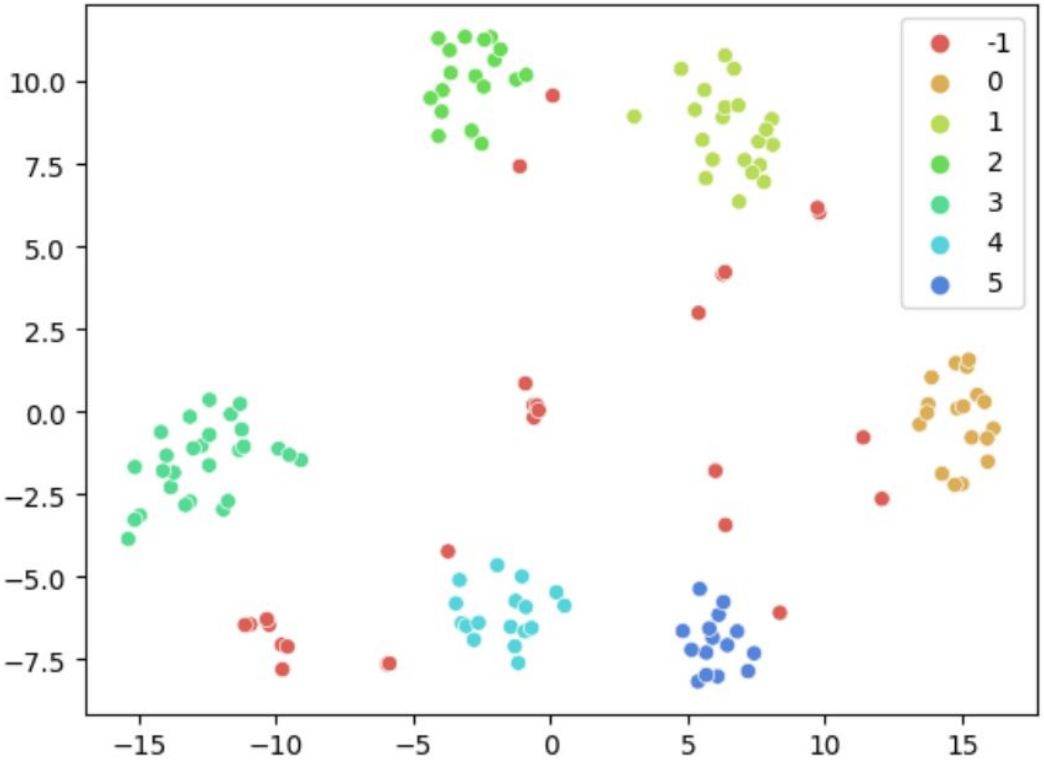
Budowane modele

- KMeans
- KMedoids
- DBSCAN
- OPTICS
- Agglomerative Clustering

Wybrany model do wdrożenia

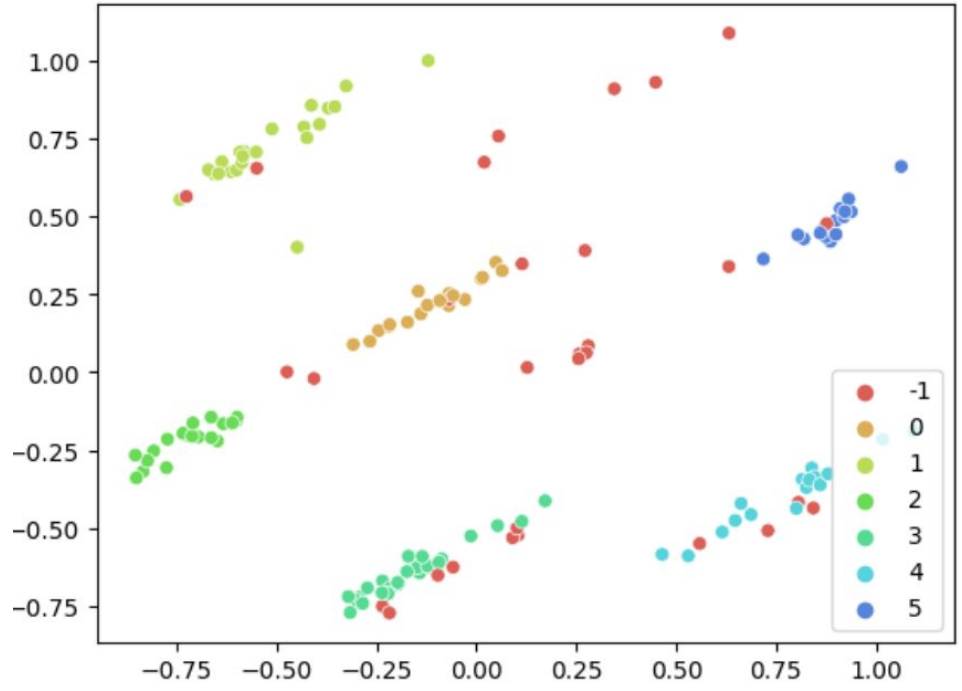
OPTICS

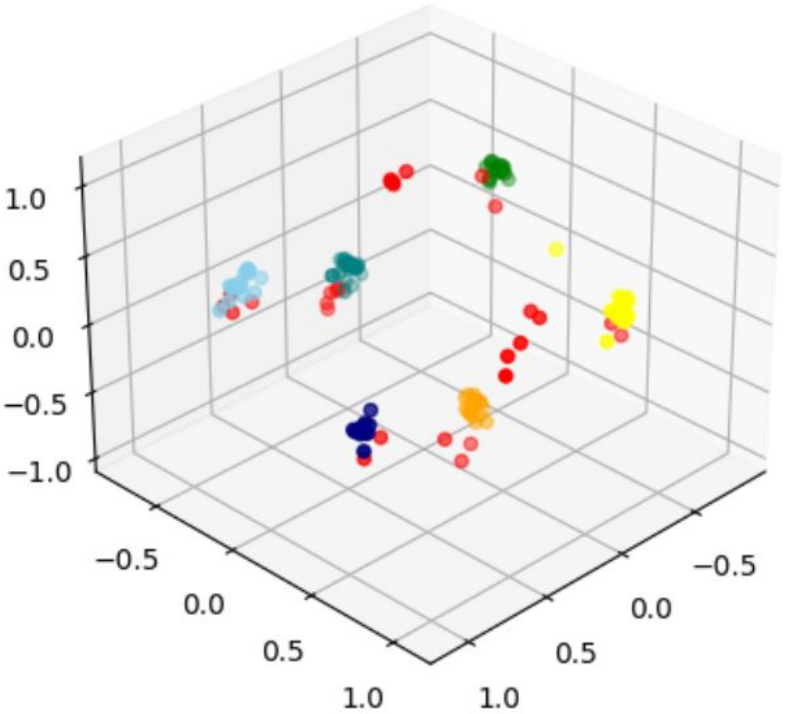
Model dokładnie grupuje obserwacje, czyli pacjentów oraz wykrywa wartości odstające w postaci szumu - klastru oznaczonego (-1)



*Wizualizacja t-SNE dla modelu OPTICS

Wizualizacja PCA dla modelu OPTICS

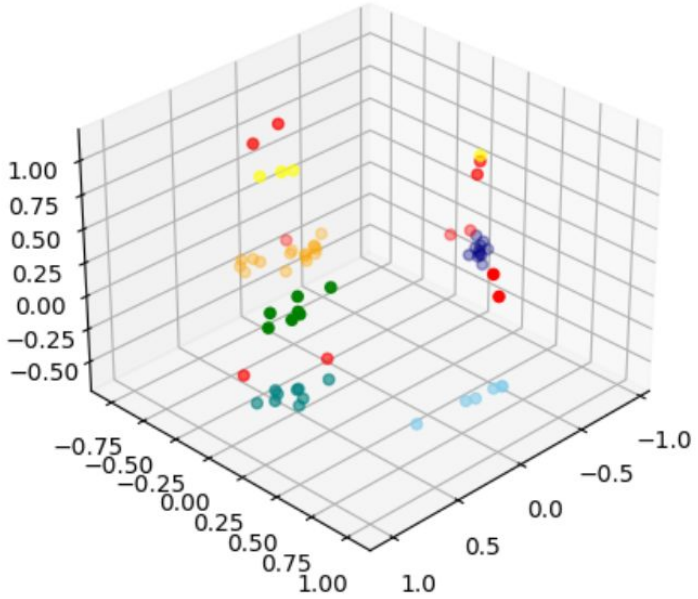




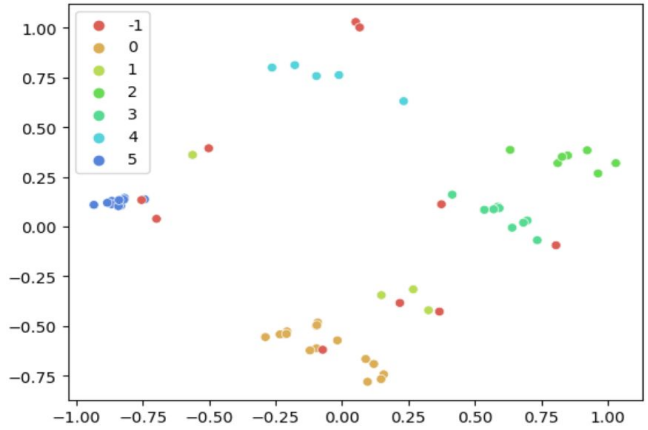
Następnie przypisujemy nowe dane testowe, których OPTICS “nie widział” do istniejących klastrów. Wykorzystujemy do tego klasyfikator K-NN

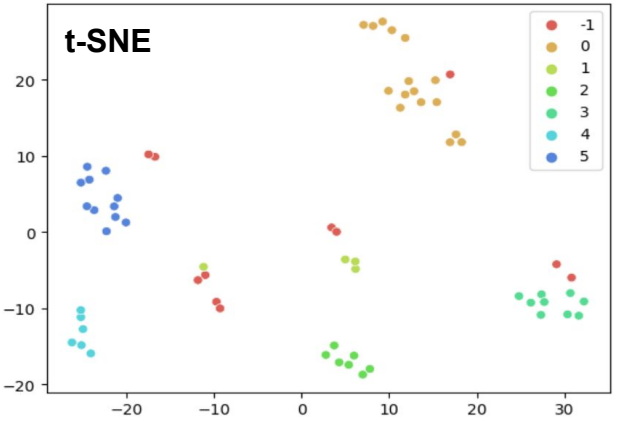
Wizualizacje klasyfikacji danych testowych

PCA 3 wymiarowe



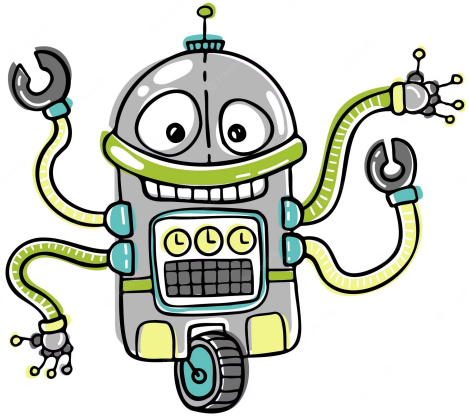
PCA 2 wymiarowe





Dane testowe również formują zwięzłe klastry mimo małej liczby obserwacji.

Interpretacja klastrów



Wartości istotności zmiennych

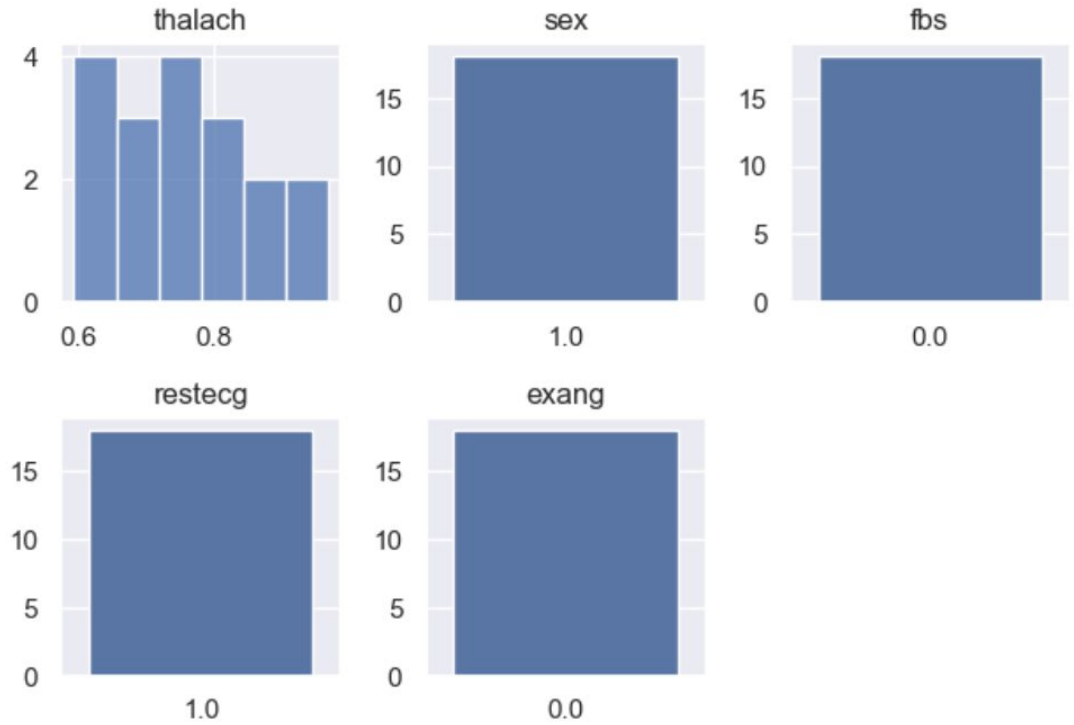
Największe znaczenie przy grupowaniu ma zmienna **wyników ekg w spoczynku** i **płeć** następnie lekko mniejszy fakt **występowania dławicy wysiłkowej**. Lekko poniżej wartości 1.0 są kolejno zmienne: **poziom cukru na czczo** oraz **maksymalne tętno**

	Feature	Importance
0	restecg	0.200034
1	sex	0.196482
2	exang	0.134793
3	fbs	0.092216
4	thalach	0.080641
5	age	0.061187
6	trestbps	0.058054
7	chol	0.056072
8	oldpeak	0.053260
9	cp	0.042715
10	slope	0.024546

Model **OPTICS** grupuje pacjentów na **6 klastrów** oraz tworzy dodatkowy klaster z pacjentami o wynikach z wartościami odstającymi

Zobaczmy charakterystyki pacjentów w poszczególnych klastrach

Klaster 0

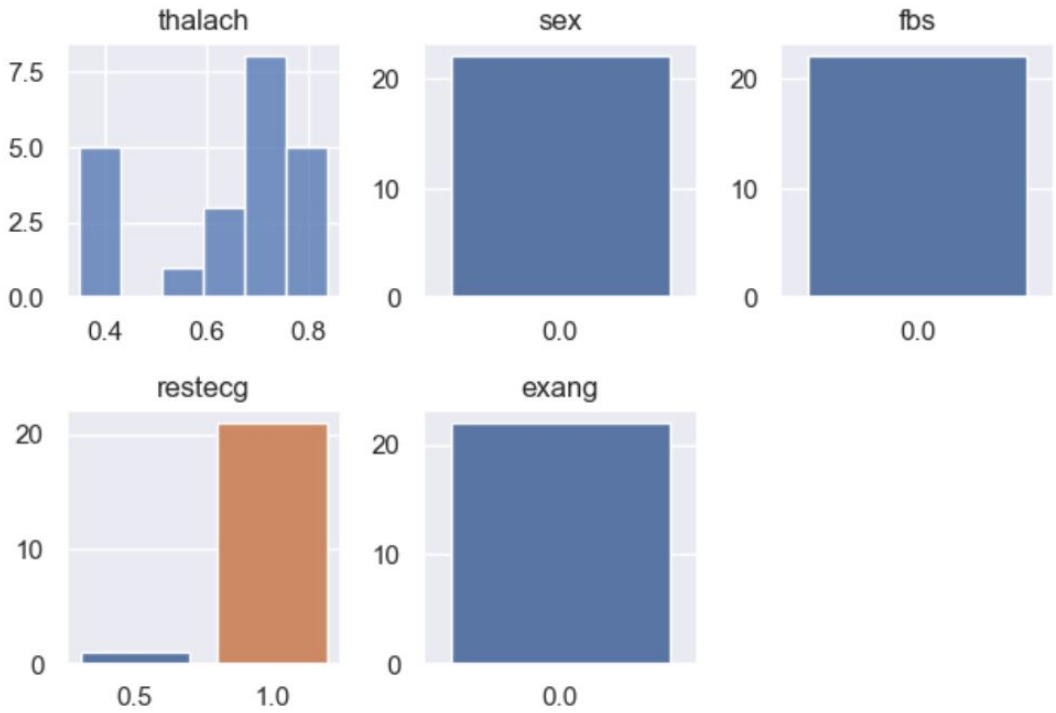


Charakterystyka pacjenta:

-mężczyzna
-poziom cukru na czczo poniżej 120 mg/dl krwi
-wyniki ekg w spoczynku w kategorii 2
-brak dławicy wysiłkowej
-mediana zmiennej maks. tętno ‘thalach’ wynosi: ~0,76 w przedziale [0,1]

Liczba pacjentów tej grupy w zbiorze treningowym: 18/148

Klaster 1

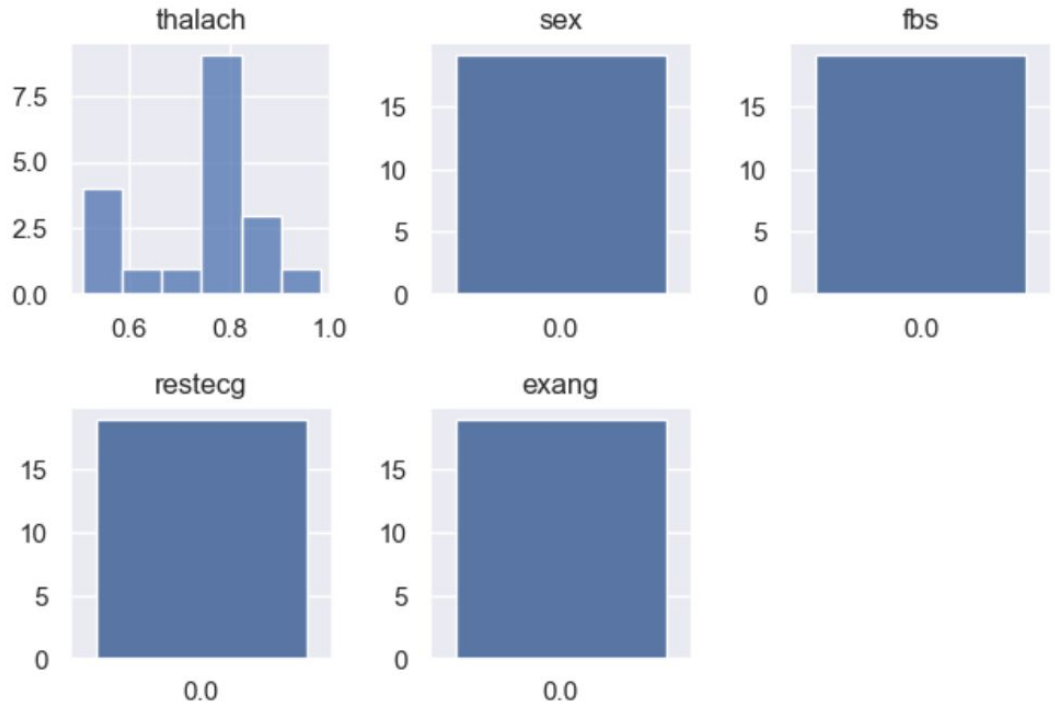


Charakterystyka pacjenta:

-kobieta
-poziom cukru na czczo poniżej 120 mg/dl krwi
-wyniki ekg w spoczynku głównie w kategorii 2, są obserwacje z kat. 1
-brak dławicy wysiłkowej
-mediana zmiennej maks. tętno ‘thalach’ wynosi: ~0,70 w przedziale [0,1]

Liczba pacjentów tej grupy w zbiorze treningowym: 22/148

Klaster 2

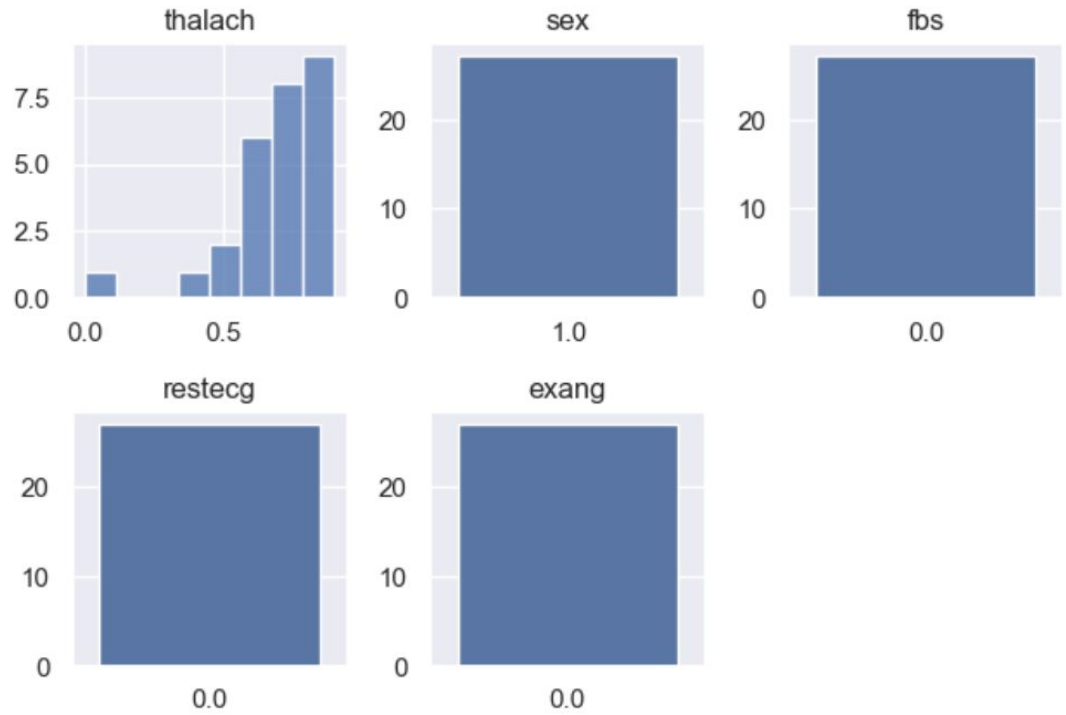


Charakterystyka pacjenta:

-kobieta
-poziom cukru na czczo poniżej 120 mg/dl krwi
-wyniki ekg w spoczynku w kat. 0
-brak dławicy wysiłkowej
-mediana zmiennej maks. tętno ‘thalach’ wynosi: ~0,76 w przedziale [0,1]

Liczba pacjentów tej grupy w zbiorze treningowym: 19/148

Klaster 3

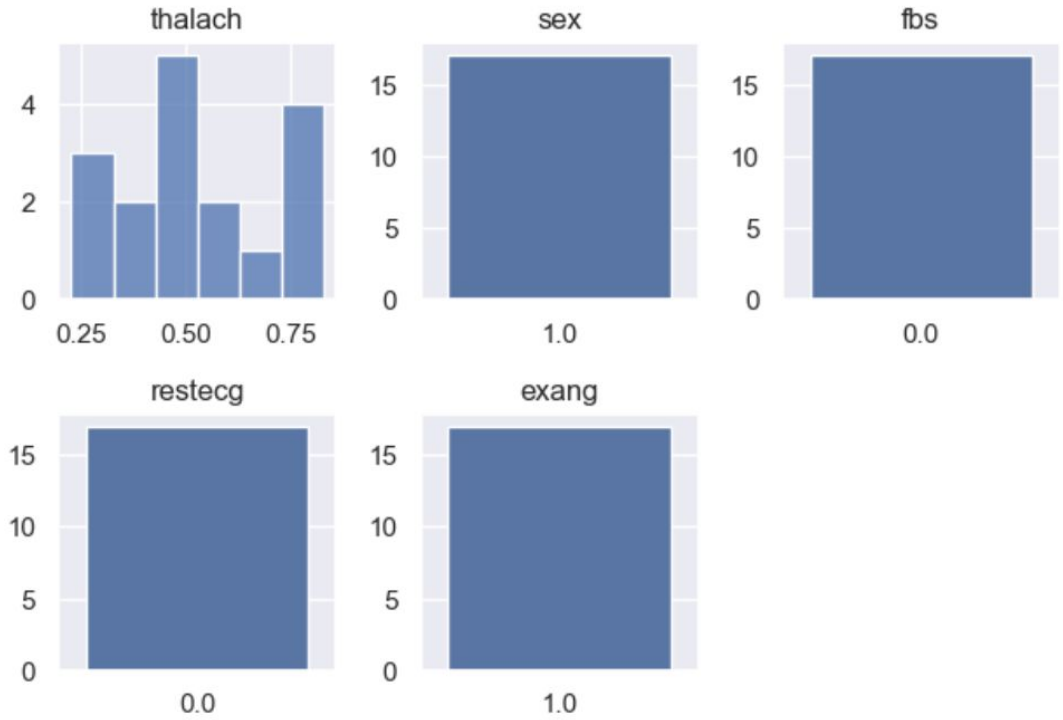


Charakterystyka pacjenta:

-mężczyzna
-poziom cukru na czczo poniżej 120 mg/dl krwi
-wyniki ekg w spoczynku w kat. 0
-brak dławicy wysiłkowej
-mediana zmiennej maks. tętno ‘thalach’ wynosi: ~0,74 w przedziale [0,1]

Liczba pacjentów tej grupy w zbiorze treningowym: 27/148

Klaster 4

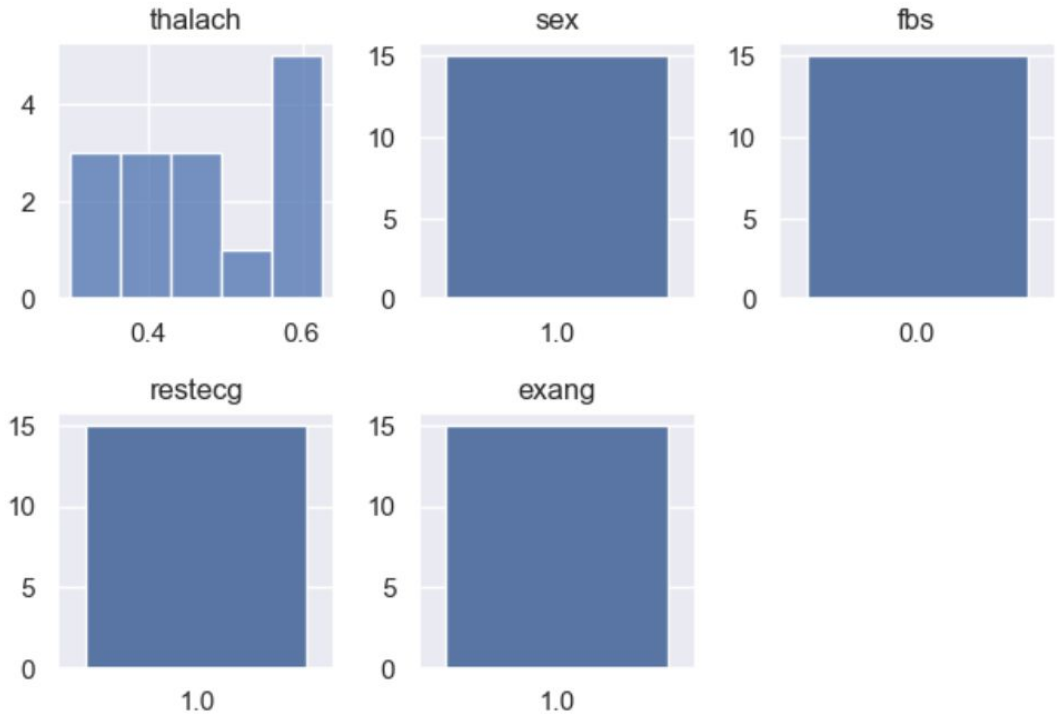


Charakterystyka pacjenta:

-mężczyzna
-poziom cukru na czczo poniżej 120 mg/dl krwi
-wyniki ekg w spoczynku w kat. 0
-obecna dławica wysiłkowa
-mediana zmiennej maks. tętno ‘thalach’ wynosi: ~0,48 w przedziale [0,1]

Liczba pacjentów tej grupy w zbiorze treningowym: 17/148

Klaster 5

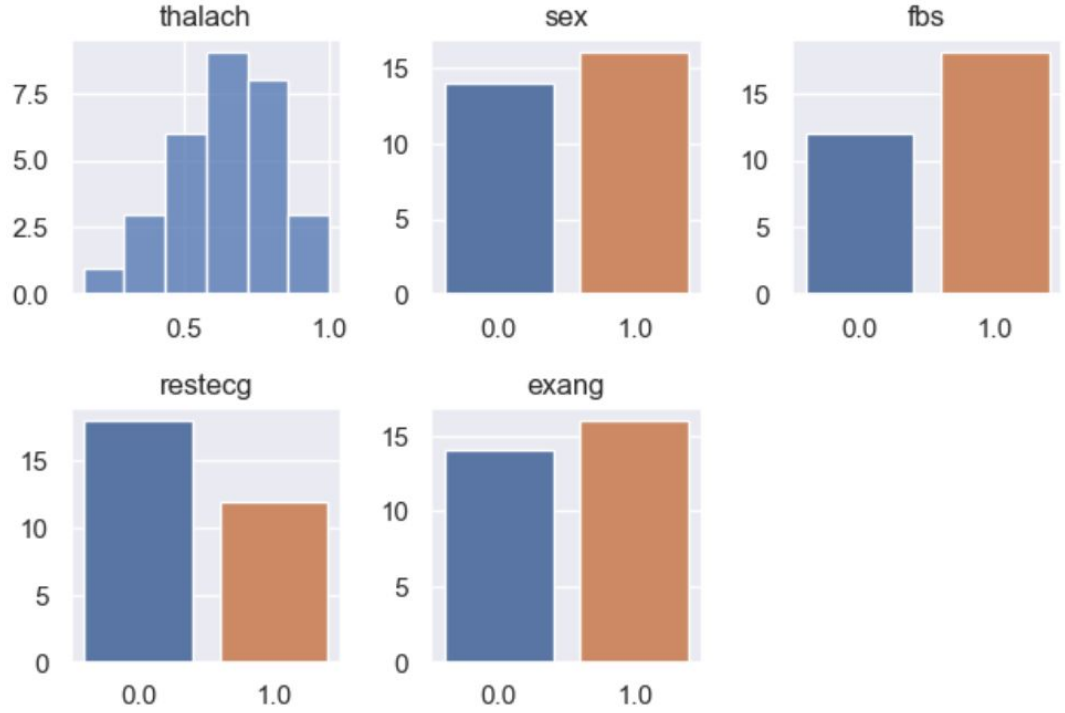


Charakterystyka pacjenta:

-mężczyzna
-poziom cukru na czczo poniżej 120 mg/dl krwi
-wyniki ekg w spoczynku w kat. 2
-obecna dławica wysiłkowa
-mediana zmiennej maks. tętno ‘thalach’ wynosi: ~0,47 w przedziale [0,1]

Liczba pacjentów tej grupy w zbiorze treningowym: 15/148

Klaster -1 pacjenci bez grupy



Charakterystyka pacjenta:

-mężczyzna lub kobieta
-w tej grupie znajdują się wszyscy pacjenci z podwyższonym poziomem cukru na czczo we krwi (hiperglikemia)
-reszta zmiennych nie daje charakterystycznych wyników

Liczba pacjentów tej grupy w zbiorze treningowym: 30/148

Uwzględniliśmy wszystkie uwagi zespołu walidacji.

Źródło danych: https://www.kaggle.com/datasets/kingabzpro/heart-disease-patients

Źródła ilustracji:

https://www.istockphoto.com/pl/ilustracje/robotic-heart?page=8

https://plus.maths.org/content/will-machine-learning-replace-mathematicians

https://pl.freepik.com/darmowe-wektory

Dziękujemy za uwagę :)