
DeepDive™

Co pamiętamy z poprzednich zajęć

DeepDive™

Jak działa PCA?

1. Standaryzacja danych
2. Obliczenie macierzy kowariancji/korelacji
3. Wyznaczenie wektorów i wartości własnych
4. Wybór składowych głównych
5. Transformacja danych

DeepDive™

Jak działa t-SNE?

1. Obliczenie podobieństw między punktami w przestrzeni wysokowymiarowej
2. Losowa inicjalizacja punktów w przestrzeni 2D lub 3D
3. Obliczenie podobieństw między punktami w niskim wymiarze
4. Minimalizacja różnicy między tymi dwoma rozkładami podobieństw
5. Przemieszczanie punktów w 2D, aż rozkłady będą jak najbardziej zbliżone

DeepDive™

Jak działa LDA?

1. **Obliczenie średnich klasowych**

   Dla każdej klasy obliczamy wektor średnich wartości cech

2. **Obliczenie macierzy rozrzutu wewnątrzklasowego**

   Reprezentuje wariancję cech w obrębie każdej klasy

3. **Obliczenie macierzy rozrzutu międzyklasowego**

   Reprezentuje wariancję cech między średnimi wartościami cech różnych klas.

4. **Rozwiązanie problemu wartości własnych**

   Pozwala znaleźć kierunki w przestrzeni cech, które maksymalizują separację między klasami

5. **Wybór głównych kierunków dyskryminacji**

   Wybieramy k-1 wektorów własnych odpowiadających największym wartościom własnym

6. **Transformacja danych**

   Rzutujemy oryginalne dane na wybrane kierunki, uzyskując nową przestrzeń o mniejszej liczbie wymiarów

DeepDive™

Czym jest regresja?

DeepDive™

Czym jest regresja?

Regresja to zbiór algorytmów statystycznych, które mają na celu dopasowanie parametrów pewnej funkcji tak, aby dana funkcja jak najlepiej odwzorowywała dane, do których dopasowanie było prowadzone.

Regresja liniowa zwykła

Najprostrzym przykładem regresji jest zwykła regresja liniowa.

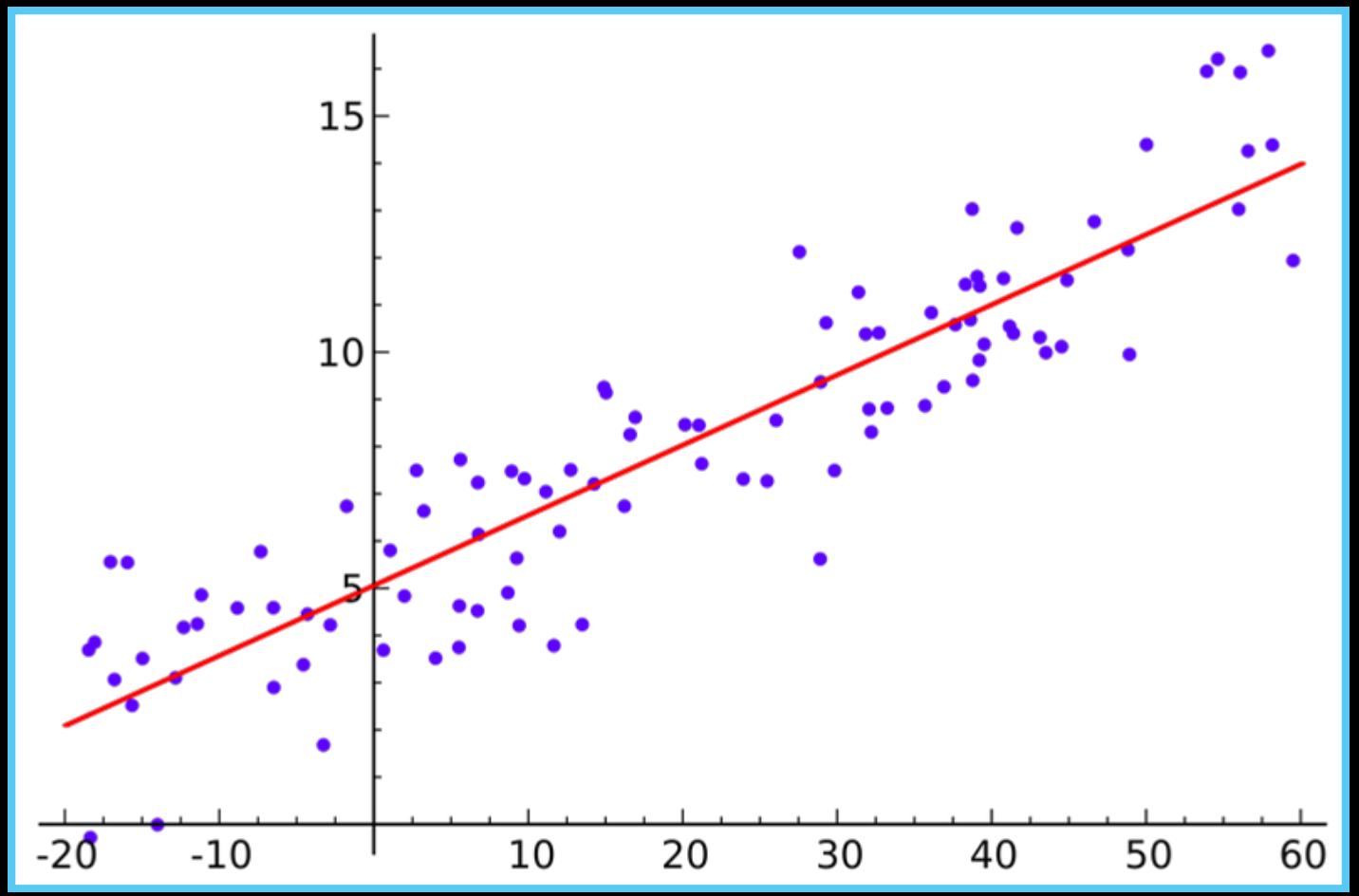
Zakłada ona, że funkcja opisująca zbiór danych ma postać:

$$y = \boldsymbol{a}x + \boldsymbol{b}$$

Gdzie rozwiązywanym problemem jest dopasowanie parametrów a i b

DeepDive™

Regresja liniowa zwykła



DeepDive™

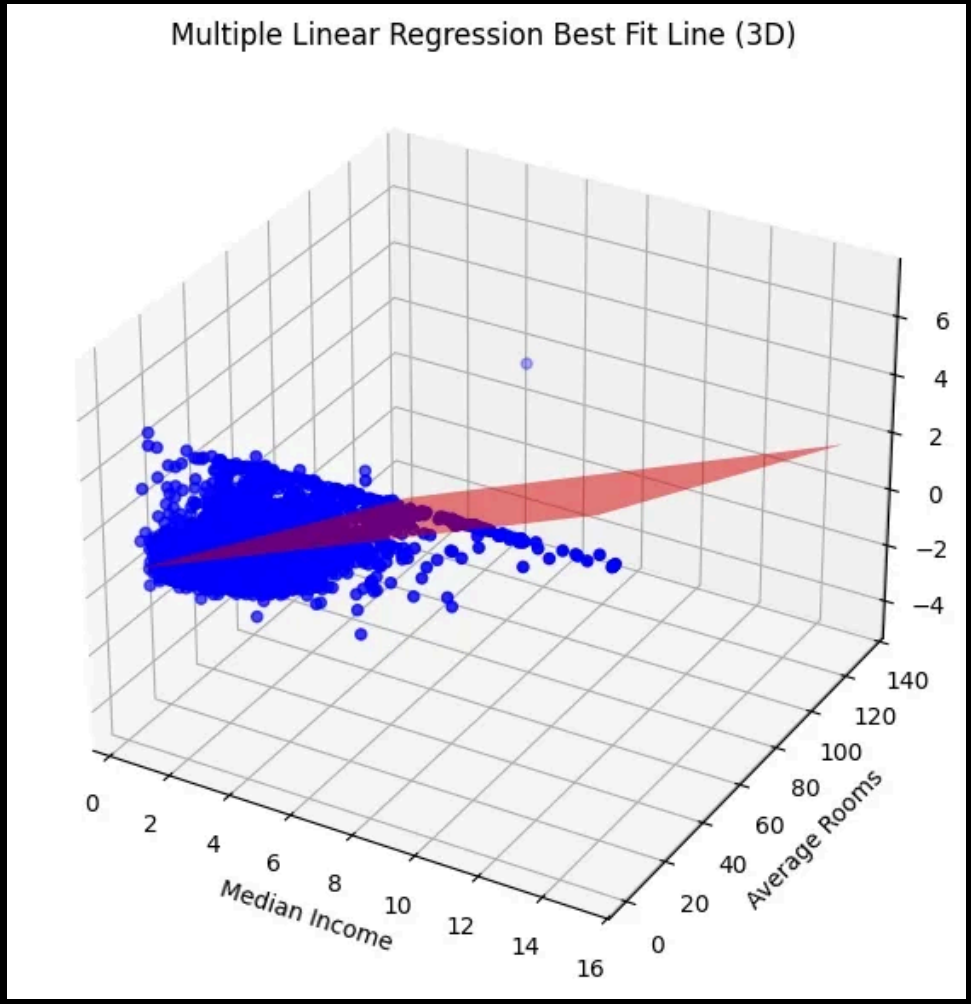
MLR - Multiple Linear Regression

Regresja MLR rozwija proste podstawy zwykłej regresji poprzez zwiększenie ilości parametrów odpowiadających za wynikową funkcję dopasowania, oraz przyjmuje większą liczbę ***zmiennych opisujących***.

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

DeepDive™

MLR – Cena domów w USA względem mediany dochodu i ilości pokoi



DeepDive™

Zmiennych ***JAKICH?***

Ano właśnie...

Zmienna opisywana (***zależna***) – to jest to coś, co **chcemy wyjaśnić**. Problem, efekt, zjawisko, ognisko dramy.

Np. „dlaczego ludzie zasypiają na wykładach?” -> zasypianie na wykładach to ***zmienna opisywana.***

DeepDive™

Zmiennych ***JAKICH?***

Zmienne opisujące (***niezależne***) – to rzeczy, które **tłumaczą, skąd się bierze zmienna opisywana.** Czyli wszystko mogące mieć wpływ na wynik.

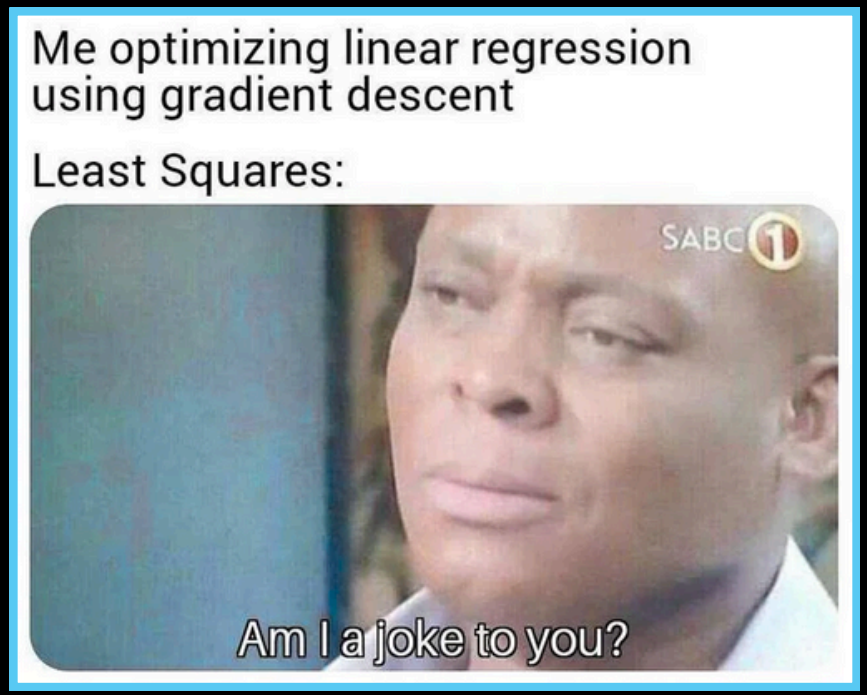
Np. nuda wykładu, czas trwania, temperatura sali, obiad z pierogami godzinę wcześniej -> ***zmienne opisujące.***

DeepDive™

OLS
(Ordinary Least Squares)

zwykła metoda najmniejszych kwadratów

celem jest minimalizacja SSE (Sum of Squared Errors)



DeepDive™

WRÓG OLS

Gdy zmienne X są bardzo silnie ze sobą skorelowane:

- wzory OLS wariują
- współczynniki Beta stają się niestabilne
- p-values rosną
- interpretacja przestaje mieć sens

Co zrobić?

Wyrzucić jedną ze skorelowanych zmiennych
albo wykonać PCA.

DeepDive™

SST – całkowita wariancja

SSR – wariancja wyjaśniona

SSE – wariancja niewyjaśniona

R2 – SSR/SST, jakość dopasowania

R2 adj. – porównywanie modeli

t-test – istoność pojedynczych zmiennych

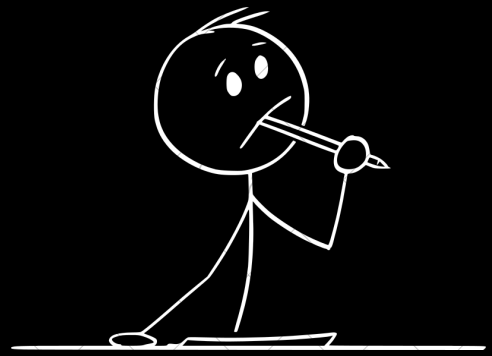
F-test – istność całego modelu

ANOVA – tabelaryczna wersja F-testu

Dobór zmiennych

dobór ręczny

forward selection



backward elimination

regularyzacja

DeepDive™

Case study

Charakterystyka grupy badawczej:

300 mężczyzn, poddano badaniu:
-pułap tlenowy
-ciśnienie skurczowe krwi [mm Hg]
-cholesterol całkowity [mg/dL]
-cholesterol HDL [mg/dL]
-trójglicerydy [mg/dL]

DeepDive™

Case study

	CIŚNIENIE	CHOLESTEROL CAŁKOWITY	CHOLESTEROL HDL	TRÓJGLICE RYDY
W NORMIE	84 %	28.33 %	67 %	44.67 %
POWYŻEJ NORMY	16 %	71.67 %	33 % (poniżej normy)	55.34 %

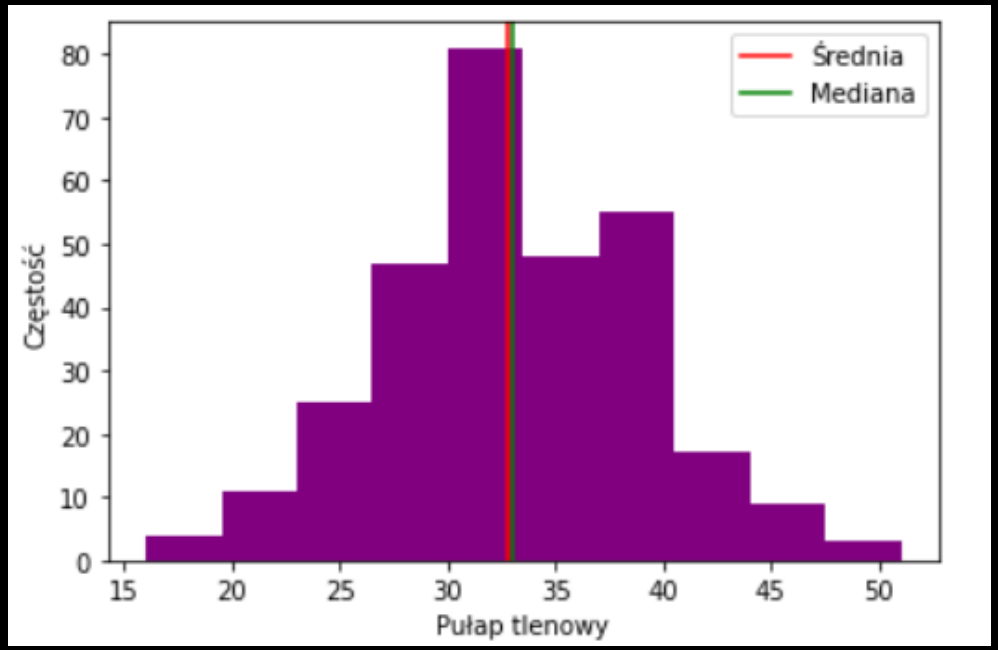
DeepDive™

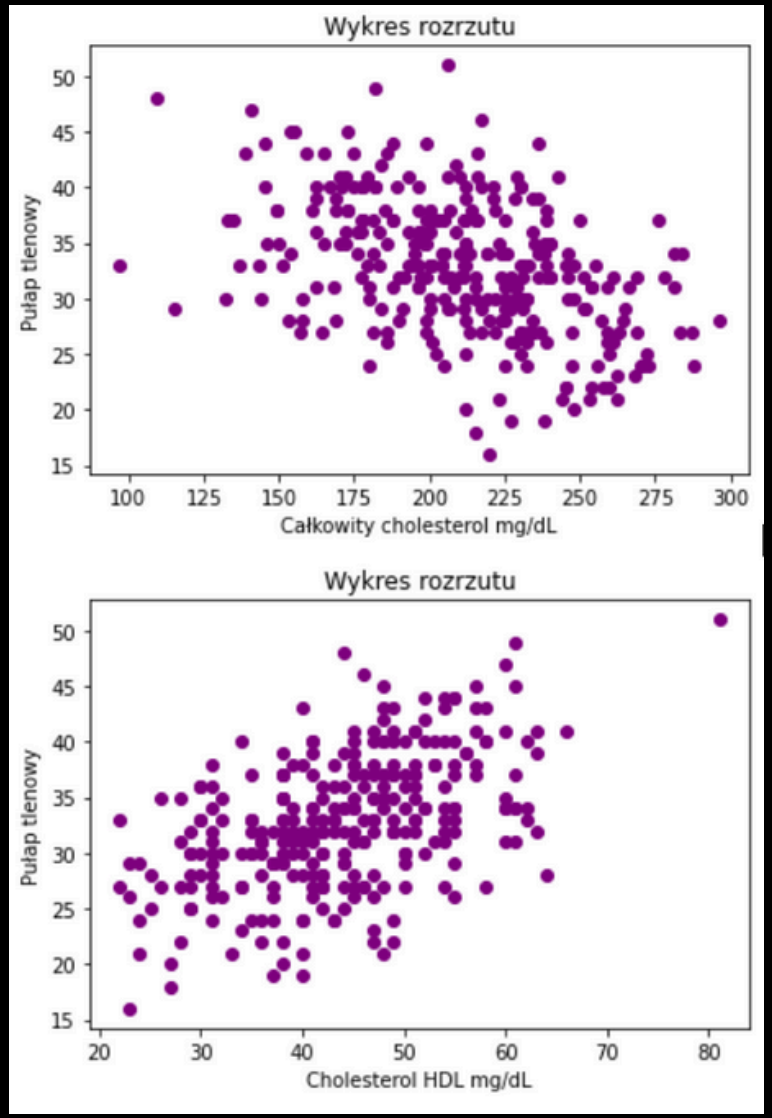
Case study

	Wartość minimalna	Wartość maksymalna	Średnia	Odchylenie standardowe	Mediana
Pułap tlenowy	16	51	32.80	6.04	33
Skurczowe ciśnienie krwi	103	162	129.096	10.59	129
Cholesterol całkowity	97	296	210.32	35.68	212
Cholesterol HDL	22	81	43.76	9.82	44
Trójglicerydy	84	239	164.46	30.49	165

DeepDive™

Case study





DeepDive™

Case study

	Pułap tlenowy	Skurczowe ciśnienie krwi	Cholesterol całkowity	Cholesterol HDL	Trójglicerydy
Pułap tlenowy	1				
Skurczowe ciśnienie krwi	-0.1298	1			
Cholesterol całkowity	-0.4495	0.5498	1		
Cholesterol HDL	0.4908	-0.3120	0.1254	1	
Trójglicerydy	0.2685	0.2949	0.3431	0.0085	1

DeepDive™

Ilość zmiennych	R-Sq	R-Sq (adj)	s	Ciśnienie	Cholesterol	HDL	Trójglicerydy
1	0.017	0.014	5.9222	x			
1	0.202	0.199	5.3354		x		
1	0.241	0.238	5.2037			x	
1	0.072	0.069	5.7533				x
2	0.222	0.217	5.2690	x	x		
2	0.242	0.236	5.2016	x		x	
2	0.120	0.114	5.6031	x			x
2	0.506	0.503	4.1968		x	x	
2	0.405	0.401	4.6087		x		x
2	0.311	0.306	4.9584			x	x
3	0.315	0.308	4.9448	x		x	x
3	0.728	0.725	3.1153		x	x	x
3	0.411	0.405	4.5836	x	x		
3	0.703	0.700	3.2540	x	x	x	
4	**0.876**	**0.874**	**2.1057**	**x**	**x**	**x**	**x**

Case study

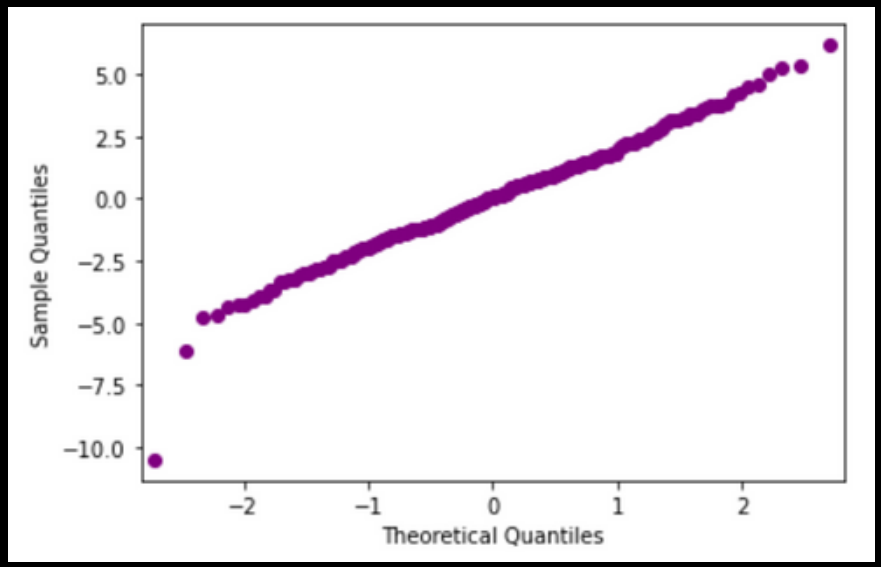
OLS Regression Results

Dep. Variable:	Pułap tlenowy	R-squared:	0.876
Model:	OLS	Adj. R-squared:	0.874
Method:	Least Squares	F-statistic:	517.8
Date:	Thu, 14 Jul 2022	Prob (F-statistic):	9.98e-132
Time:	07:33:32	Log-Likelihood:	-645.92
No. Observations:	299	AIC:	1302.
Df Residuals:	294	BIC:	1320.
Df Model:	4		
Covariance Type:	nonrobust		

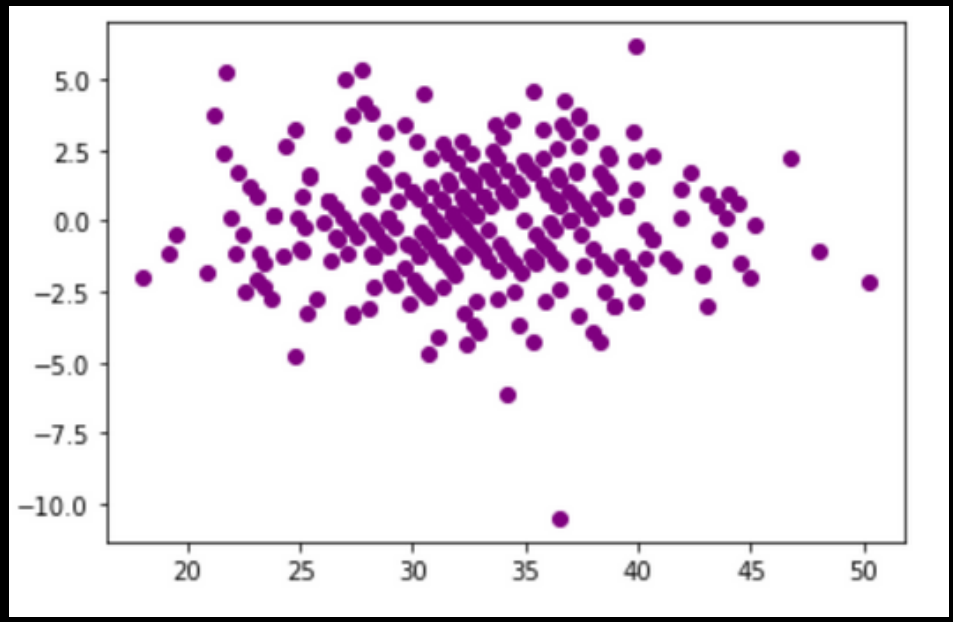
	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.9850	1.946	-3.075	0.002	-9.816	-2.154
Skurczowe ciśnienie krwi [mm Hg]	0.2946	0.016	18.694	0.000	0.264	0.326
Całkowity cholesterol mg/dL	-0.1647	0.005	-36.431	0.000	-0.174	-0.156
Cholesterol HDL mg/dL	0.4809	0.015	33.151	0.000	0.452	0.509
Trójglicerydy mg/dL	0.0872	0.004	20.200	0.000	0.079	0.096

Omnibus:	19.190	Durbin-Watson:	1.946
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.947
Skew:	-0.306	Prob(JB):	1.28e-09
Kurtosis:	4.706	Cond. No.	4.80e+03

Case study



wykresy reszt



DeepDive™

Statystyki

Klasyczne testy statystyczne jako szczególne przypadki regresji liniowej		
Nazwa zwyczajowa	**Równoważny model liniowy**	**Opis słowny**
test *t* Studenta dla jednej próby	$y = \beta_0 + \epsilon$	Czy średnia (lub mediana) obserwacji jest ich dobrym predyktorem?
test Wilcoxona dla jednej próby	$\operatorname{ranga}^{+}_{-}(y) = \beta_0 + \epsilon$	
test *t* Studenta dla par obserwacji	$y_2 - y_1 = \beta_0 + \epsilon$	Czy średnia (lub mediana) różnic obserwacji jest ich dobrym predyktorem?
test Wilcoxona dla par obserwacji	$\operatorname{ranga}^{+}_{-}(y_2 - y_1) = \beta_0 + \epsilon$	
korelacja *r* Pearsona	$y = \beta_0 + \beta_1 x + \epsilon$	Czy model liniowy jest dobrym predyktorem obserwacji (lub ich rang)?
korelacja Spearmana	$\operatorname{ranga}(y) = \beta_0 + \beta_1 \operatorname{ranga}(x) + \epsilon$	
test *t* Studenta dla dwóch prób	$y = \beta_0 + \beta_1 D + \epsilon$	Czy średnie grup są dobrym predyktorem obserwacji (lub ich rang)?
test Manna-Whitneya	$\operatorname{ranga}^{+}_{-}(y) = \beta_0 + \beta_1 D + \epsilon$	
jednoczynnikowa ANOVA	$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_n D_n + \epsilon$	
test Kruskala-Wallisa	$\operatorname{ranga}^{+}_{-}(y) = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_n D_n + \epsilon$	
jednoczynnikowa ANCOVA	$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_n D_n + \beta_x x + \epsilon$	Czy średnie grup oraz ich liniowy model są dobrym predyktorem obserwacji (lub ich rang)?
dwuczynnikowa ANOVA	$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_n D_n + \beta_o E_1 + \beta_p E_2 + \ldots + \beta_r E_m + \beta_s D_1 E_1 + \beta_t D_1 E_2 + \ldots + \beta_u D_n E_m + \epsilon$	Czy średnie grup oraz ich iloczynów są dobrym predyktorem obserwacji?

DeepDive™

A kiedy mamy problem nieliniowy?

Problemy nieliniowe bywają często wielowymiarowe. Idealną metodą regresji w takim przypadku jest na przykład Proces Gaussiański, będący podejściem Bayesowskim.

Proces Gaussiański jest podejściem ***nieparametrycznym***, czyli takim, które nie zakłada z góry kształtu wynikowej funkcji czy zależności między zmiennymi, a.k.a “Pokaż dane, ja w nie spojrze i coś wykminie” lub FA-FO.

MLR ze statsmodels

```
import statsmodels.api as sm
import pandas as pd

y = df["y"]
X = df[["x1", "x2", "x3"]]

X = sm.add_constant(X)

model = sm.OLS(y, X).fit()

print(model.summary())
```

DeepDive™

Klasyfikacja Danych

DeepDive™

Czym jest klasyfikacja?

DeepDive™

Czym jest klasyfikacja?

Jak sama nazwa wskazuje, jest to zbiór algorytmów przeznaczonych w celu podziału zbioru danych na poszczególne klasy względem zmiennych zależnych i niezależnych opisujących dane klasy.

DeepDive™

Maszyna wektorów nośnych

SUPPORT VECTOR MACHINE (SVM)

Głównym celem jest **znalezienie hiperpłaszczyzny**,
która najlepiej oddziela różne klasy w danych,
maksymalizując margines – odległość między najbliższymi punktami
danych (tzw. wektorami nośnymi) a hiperpłaszczyzną.

SVM jest szczególnie efektywny w przypadkach,
gdy dane są nieliniowo rozdzielne,
ponieważ umożliwia użycie tzw. **jądra (kernel)**,
które przekształca dane do wyższych wymiarów,
w których stają się one liniowo rozdzielne.

DeepDive™

Rodzaje funkcji jądra

liniowe (Linear)

$$K(x, y) = x^T y$$

wielomianowe (Polynomial)

$$K(x, y) = (x^T y + c)^d$$

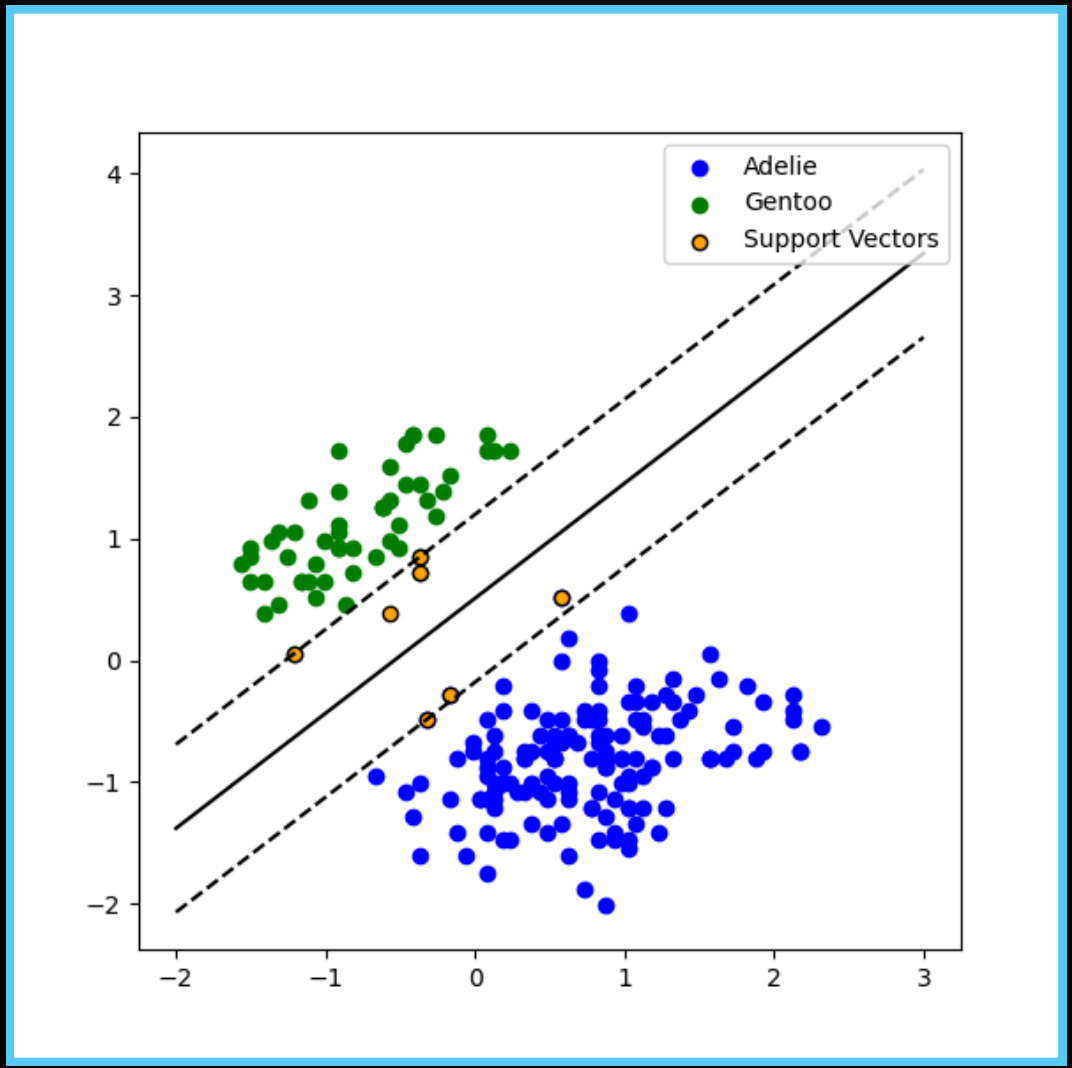
Gdzie c to stała,
a d to stopień wielomianu

radialne (RBF - Radial Basis Function)

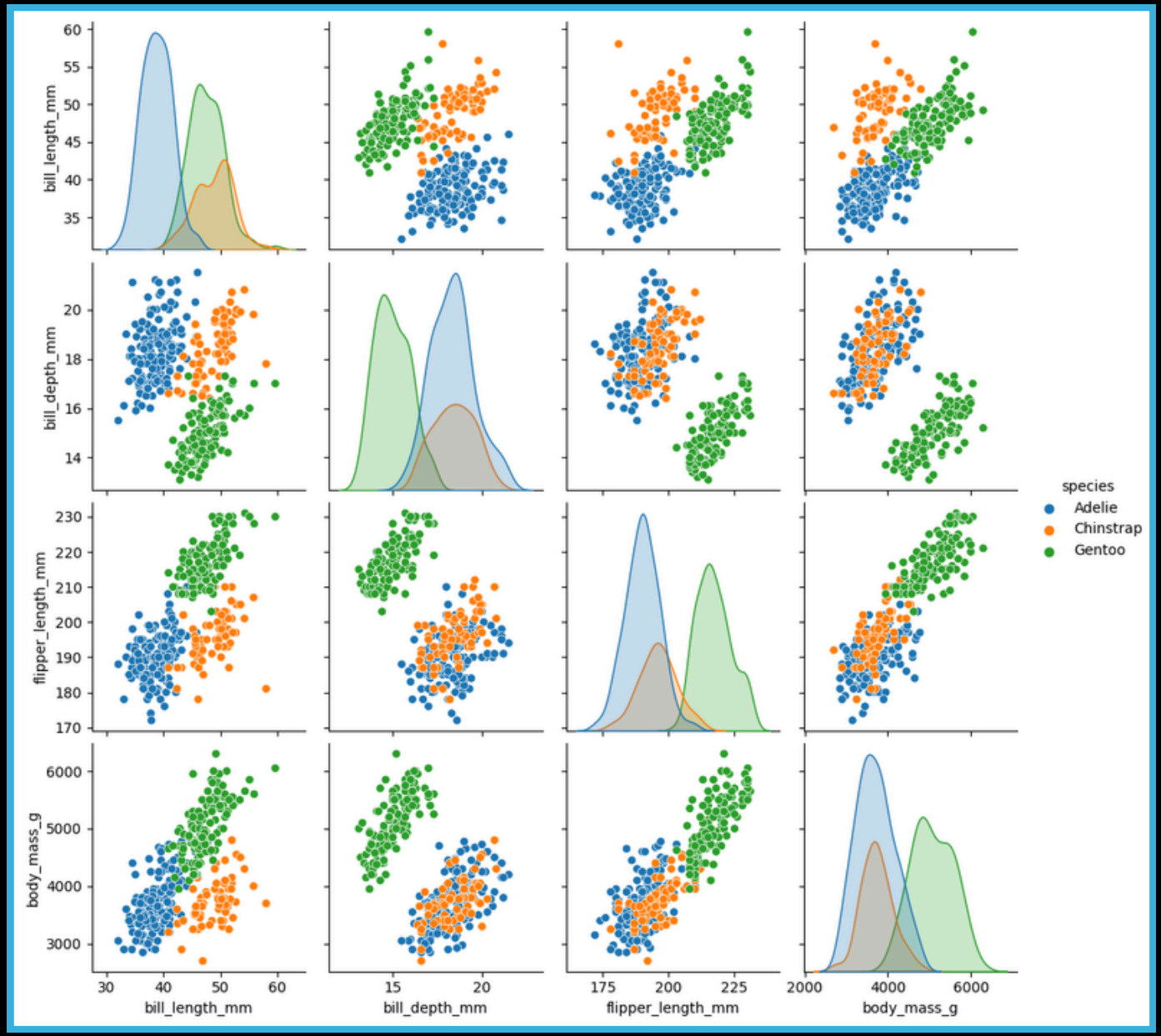
$$K(x, y) = exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

sigmoidalne (Sigmoid)

DeepDive™



DeepDive™

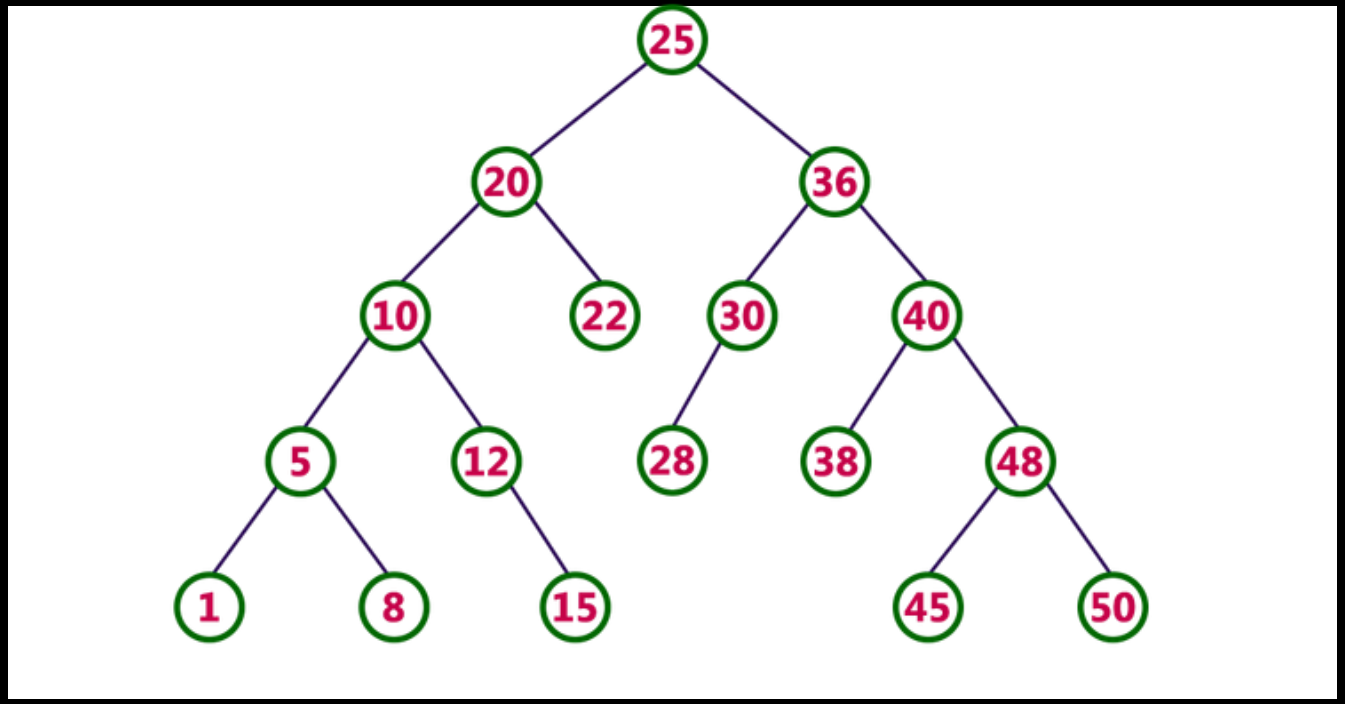

bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g
species
Adelie
Chinstrap
Gentoo

Drzewa decyzyjne i Random Forest

Drzewa decysyjne są klasyfikatorami opartymi na drzewach binarnych, dzie decyzja o przypisaniu obiektu do klasy podejmowana jest na zasadzie pytań do analizowanych cech.

DeepDive™

Drzewa decyzyjne i Random Forest



DeepDive™

Drzewa decyzyjne i Random Forest

Pojedyncze drzewa są jednak bardzo podatne na przetrenowanie, używa się więc zbioru drzew, które wybierają losowe cechy i na ich podstawie przewidują przynależność do danej klasy. Decyzja jest podejmowana przez zestawienie prawdopodobieństw przynależności z całego zbioru przewidywań.

DeepDive™


DeepDive™