Matematyka stosowana to nie kreda i tablica!

Mateusz Krukowski

Analiza danych, poziom: szkoła

Zadanie 31.

Właściciel pewnej apteki przeanalizował dane dotyczące liczby obsługiwanych klientów z 30 kolejnych dni. Przyjmijmy, że liczbę L obsługiwanych klientów n-tego dnia opisuje funkcja

$$L(n) = -n^2 + 22n + 279$$

gdzie n jest liczbą naturalną spełniającą warunki $n \geq 1$ i $n \leq 30$.

Zadanie 31.1. (0–1)

Oceń prawdziwość poniższych stwierdzeń. Wybierz P, jeśli stwierdzenie jest prawdziwe, albo F – jeśli jest fałszywe.

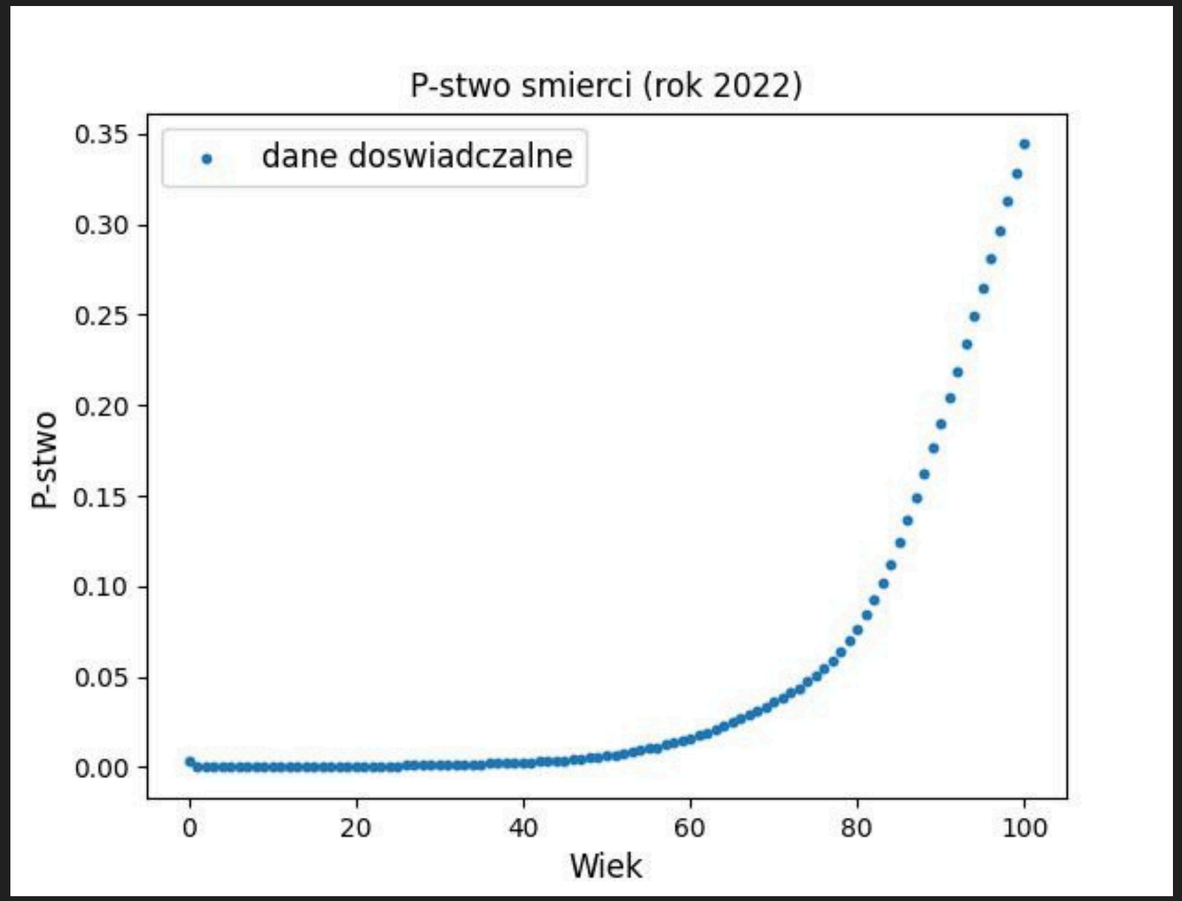
Łączna liczba klientów obsłużonych w czasie wszystkich analizowanych dni jest równa $L(30)$.	**P**	**F**
W trzecim dniu analizowanego okresu obsłużono 336 klientów.	**P**	**F**

Analiza danych

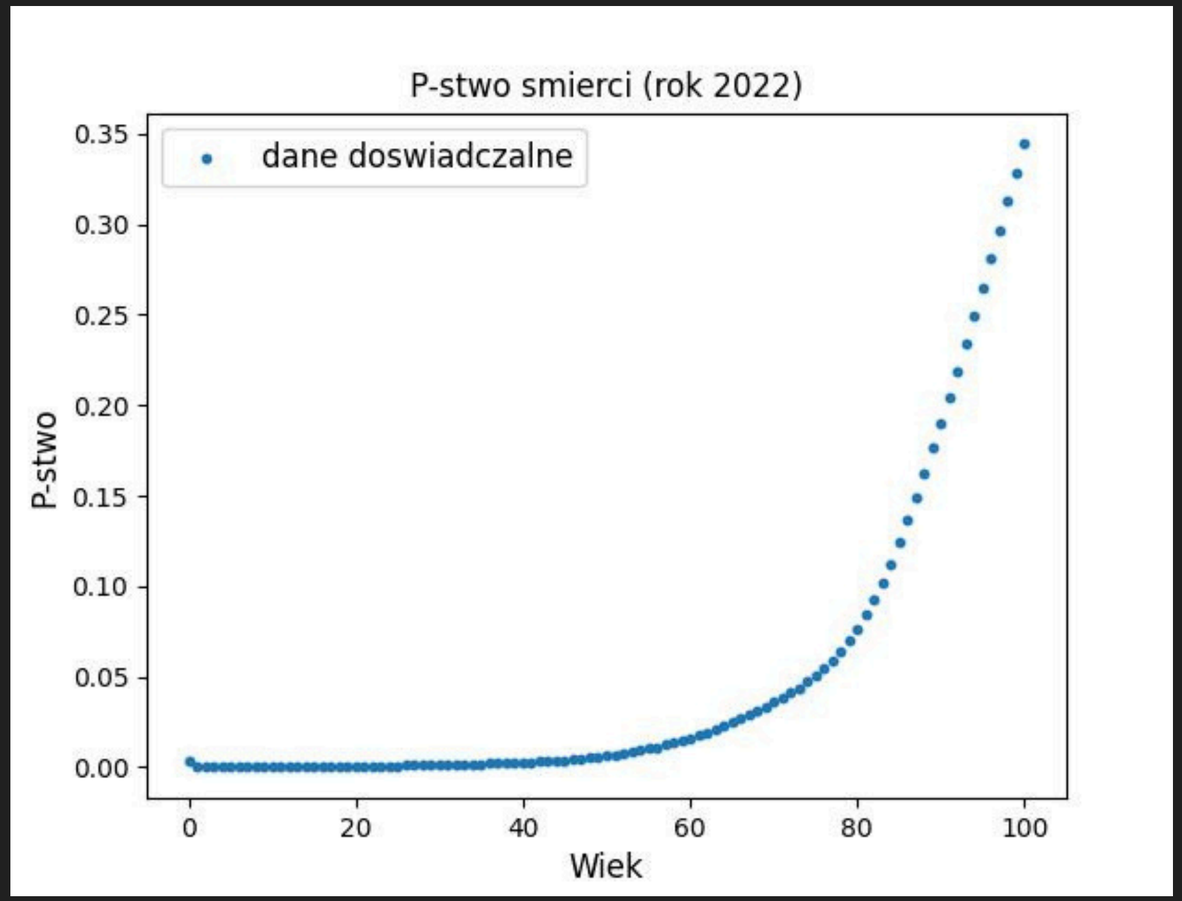
przykład: polisy ubezpieczeniowe

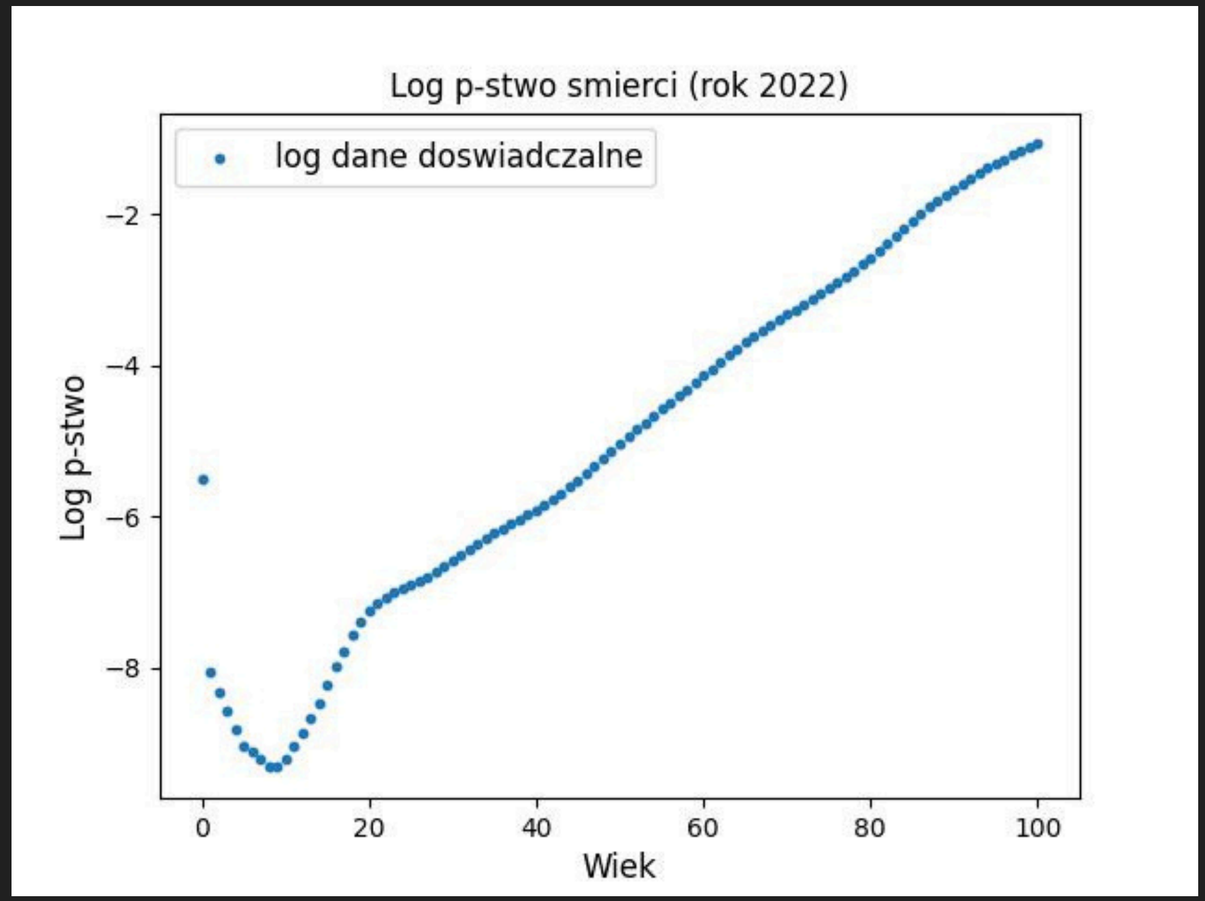
	A	B	C	D	E	F	G	H
1	**Tablica trwania życia 2022**							
2								
3	Płeć 1-mężcz 2-kobiety	Wiek	Liczba dożywa-jących	Prawdo-podobień-stwo zgonu	Liczba zmarłych	Ludność stacjonar-na	Ludność stacjonar-na skumulo-wana	Przeciętne dalsze trwanie życia
4		x	lx	qx	dx	Lx	Tx	ex
5	1	0	100000	0,00408	409	99632	7341781	73,42
6	1	1	99591	0,00032	32	99575	7242150	72,72
7	1	2	99559	0,00024	24	99547	7142575	71,74
8	1	3	99535	0,00019	19	99526	7043028	70,76
9	1	4	99516	0,00015	15	99509	6943502	69,77
10	1	5	99501	0,00012	12	99495	6843994	68,78
11	1	6	99489	0,00011	11	99484	6744499	67,79
12	1	7	99478	0,00010	9	99474	6645015	66,80
13	1	8	99469	0,00009	9	99465	6545542	65,80
14	1	9	99460	0,00009	10	99455	6446077	64,81
15	1	10	99450	0,00010	10	99445	6346622	63,82
16	1	11	99440	0,00012	12	99434	6247177	62,82
17	1	12	99428	0,00014	13	99422	6147743	61,83
18	1	13	99415	0,00017	17	99407	6048322	60,84
19	1	14	99398	0,00021	21	99388	5948915	59,85
20	1	15	99377	0,00027	27	99364	5849528	58,86

Polisy ubezpieczeniowe

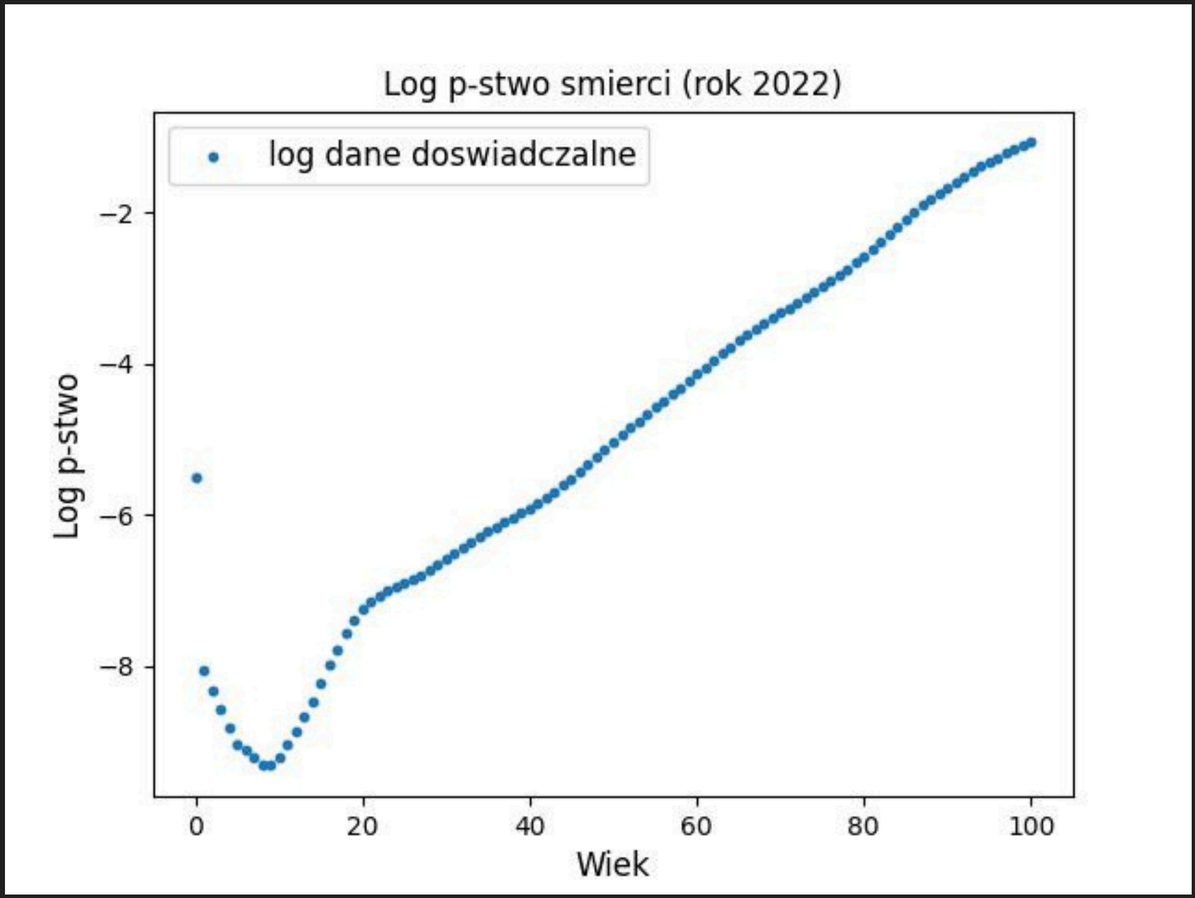


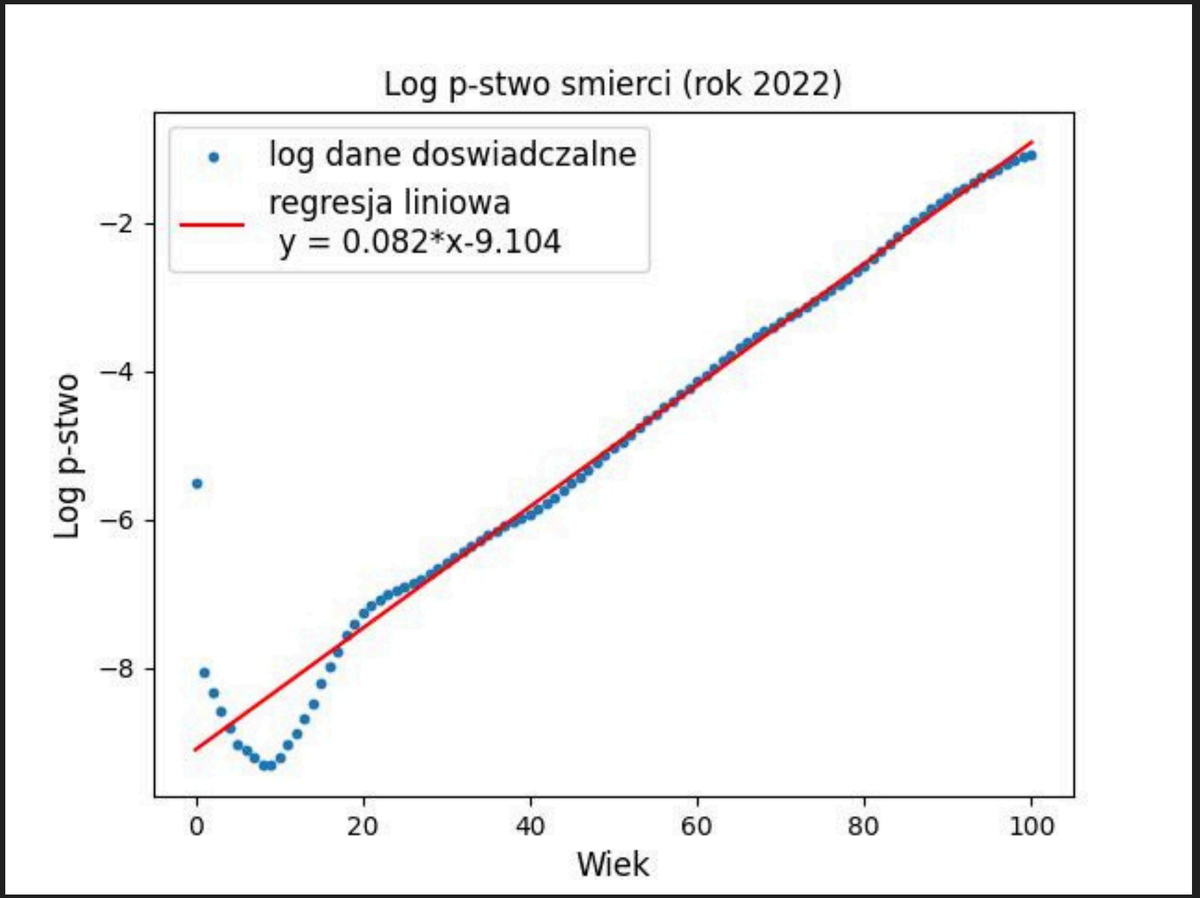
Polisy ubezpieczeniowe



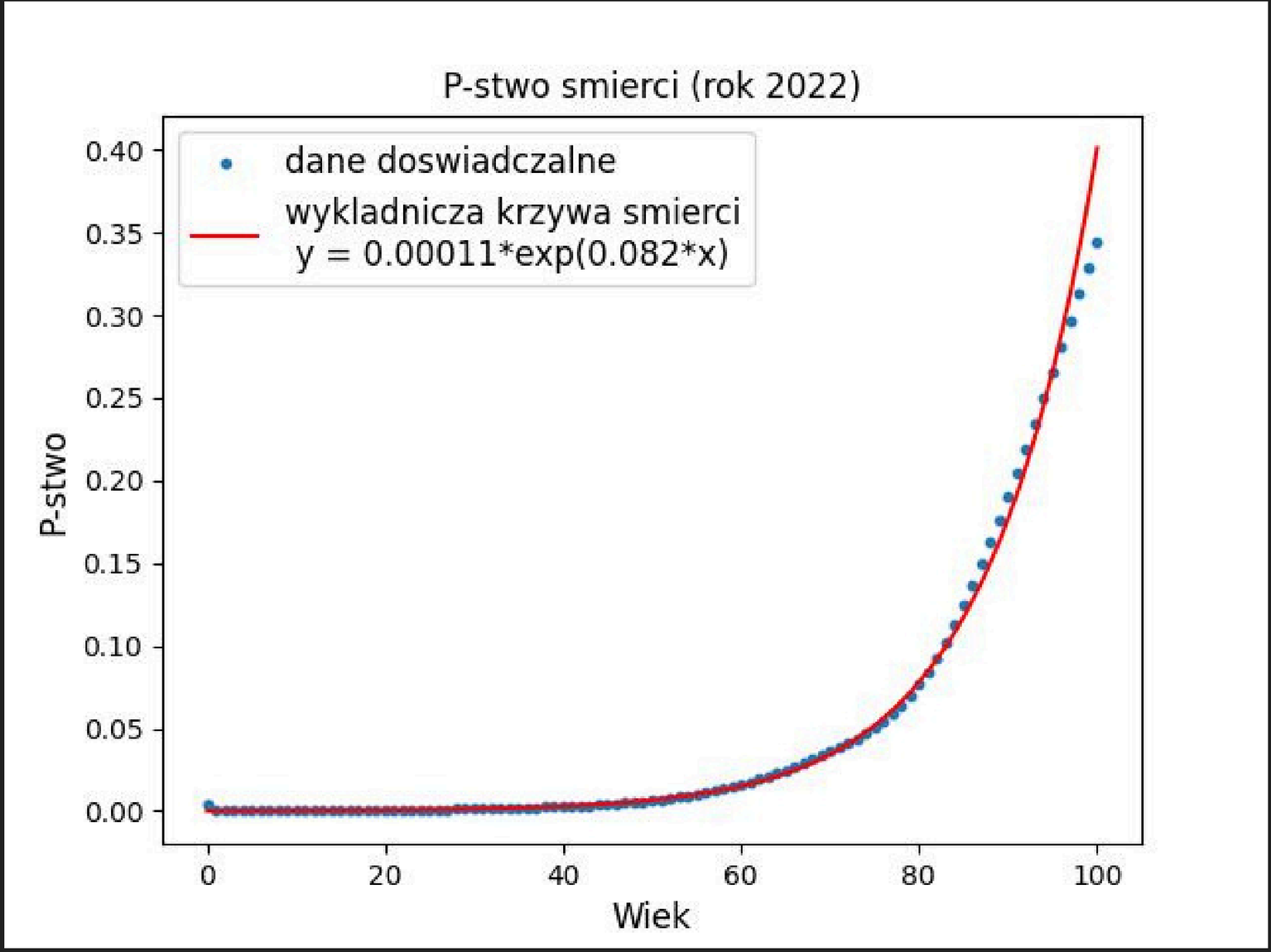


Polisy ubezpieczeniowe





Polisy ubezpieczeniowe



Problem klasyfikacji

CEL: przydzielenie etykiety do obiektu na podstawie cech opisujących ten obiekt.

Problem klasyfikacji

CEL: przydzielenie etykiety do obiektu na podstawie cech opisujących ten obiekt.

naiwny przykład: jeśli ktoś waży 95 kg i mierzy 1.87 m to spodziewamy się, że jest to...

Problem klasyfikacji

CEL: przydzielenie etykiety do obiektu na podstawie cech opisujących ten obiekt.

naiwny przykład: jeśli ktoś waży 95 kg i mierzy 1.87 m to spodziewamy się, że jest to... mężczyzna.

	Male			Female		
Country	Average height	Weight	BMI	Average height	Weight	BMI
Croatia	1.81 m	92.3 kg	28.3	1.67 m	75.3 kg	27.1
Serbia	1.80 m	89.2 kg	27.4	1.68 m	72.3 kg	25.6
Sweden	1.80 m	86.2 kg	26.5	1.67 m	69.9 kg	25.2
Norway	1.80 m	88.1 kg	27.1	1.66 m	71.7 kg	25.9
Lithuania	1.80 m	90.0 kg	27.7	1.67 m	74.2 kg	26.5
Poland	1.80 m	91.6 kg	28.2	1.65 m	72.6 kg	26.5
Ukraine	1.80 m	86.3 kg	26.6	1.66 m	74.3 kg	26.9
Finland	1.80 m	87.3 kg	26.9	1.66 m	72.9 kg	26.4
Latvia	1.80 m	89.2 kg	27.5	1.68 m	74.5 kg	26.4
Germany	1.80 m	87.8 kg	27.1	1.66 m	71.4 kg	25.9
Dominica	1.80 m	81.7 kg	25.3	1.67 m	82.4 kg	29.7
Belgium	1.79 m	84.9 kg	26.5	1.64 m	69.9 kg	26.1

Przykłady klasyfikacji (scikit.datasets)

scikit learn Install User Guide API Examples Community More

world machine learning tasks.

7.1.1. Iris plants dataset

Data Set Characteristics:

Number of Instances:	150 (50 in each of three classes)
Number of Attributes:	4 numeric, predictive attributes and the class

Attribute Information:	• sepal length in cm
	• sepal width in cm
	• petal length in cm
	• petal width in cm
	• **class:**
	  ◦ Iris-Setosa
	  ◦ Iris-Versicolour
	  ◦ Iris-Virginica

Przykłady klasyfikacji (scikit.datasets)

scikit learn Install User Guide API Examples Community More

7.1.6. Breast cancer wisconsin (diagnostic) dataset

Data Set Characteristics:

Number of Instances:	569
Number of Attributes:	30 numeric, predictive attributes and the class
Attribute Information:	• radius (mean of distances from center to points on the perimeter) • texture (standard deviation of gray-scale values) • perimeter • area • smoothness (local variation in radius lengths) • compactness (perimeter^2 / area - 1.0) • concavity (severity of concave portions of the contour) • concave points (number of concave portions of the contour) • symmetry

Przykłady klasyfikacji (UCI ML Repository)

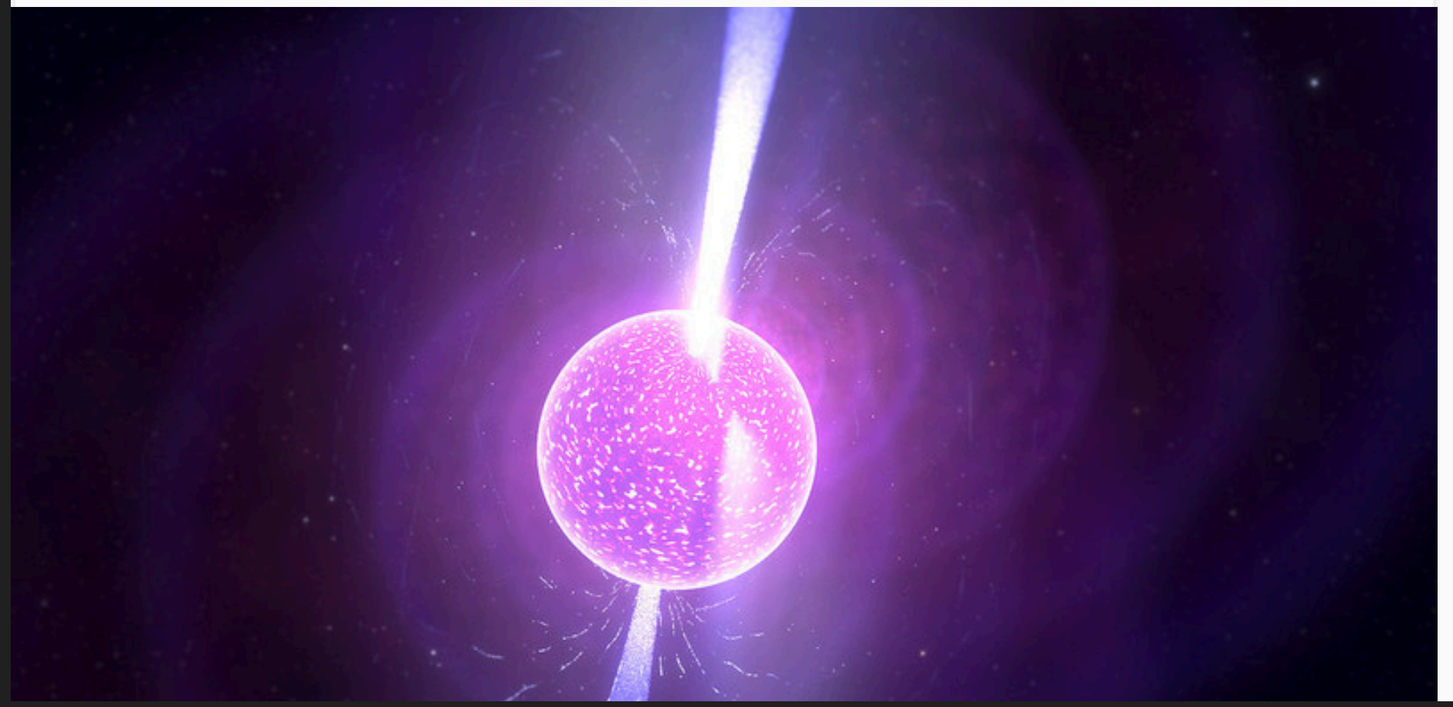
HTRU2

Donated on 2/13/2017

Pulsar candidates collected during the HTRU survey. Pulsars are a type of star, of considerable scientific interest. Candidates must be classified in to pulsar and non-pulsar classes to aid discovery.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Physics and Chemistry	Classification, Clustering

Feature Type	# Instances	# Features
Real	17898	8



DOWNLOAD (1.5 MB)

IMPORT IN PYTHON

CITE

1 citations

13395 views

Citations/Acknowledgements

If you use this dataset, please cite:

If you use the dataset in your work, please cite us using the following paper:

R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsa...

Creators

Robert Lyon

DOI

10.24432/C5DK6R

License

This dataset is licensed under a Creative

Przykłady klasyfikacji (UCI ML Repository)

HIGGS

Donated on 2/11/2014

This is a classification problem to distinguish between a signal process which produces Higgs bosons and a background process which does not.

Dataset Characteristics
-

Subject Area
Physics and Chemistry

Associated Tasks
Classification

Feature Type
Real

 # Instances
11000000

 # Features
-

DOWNLOAD (2.6 GB)

CITE

1 citations

21851 views

Creators

Daniel Whiteson

DOI

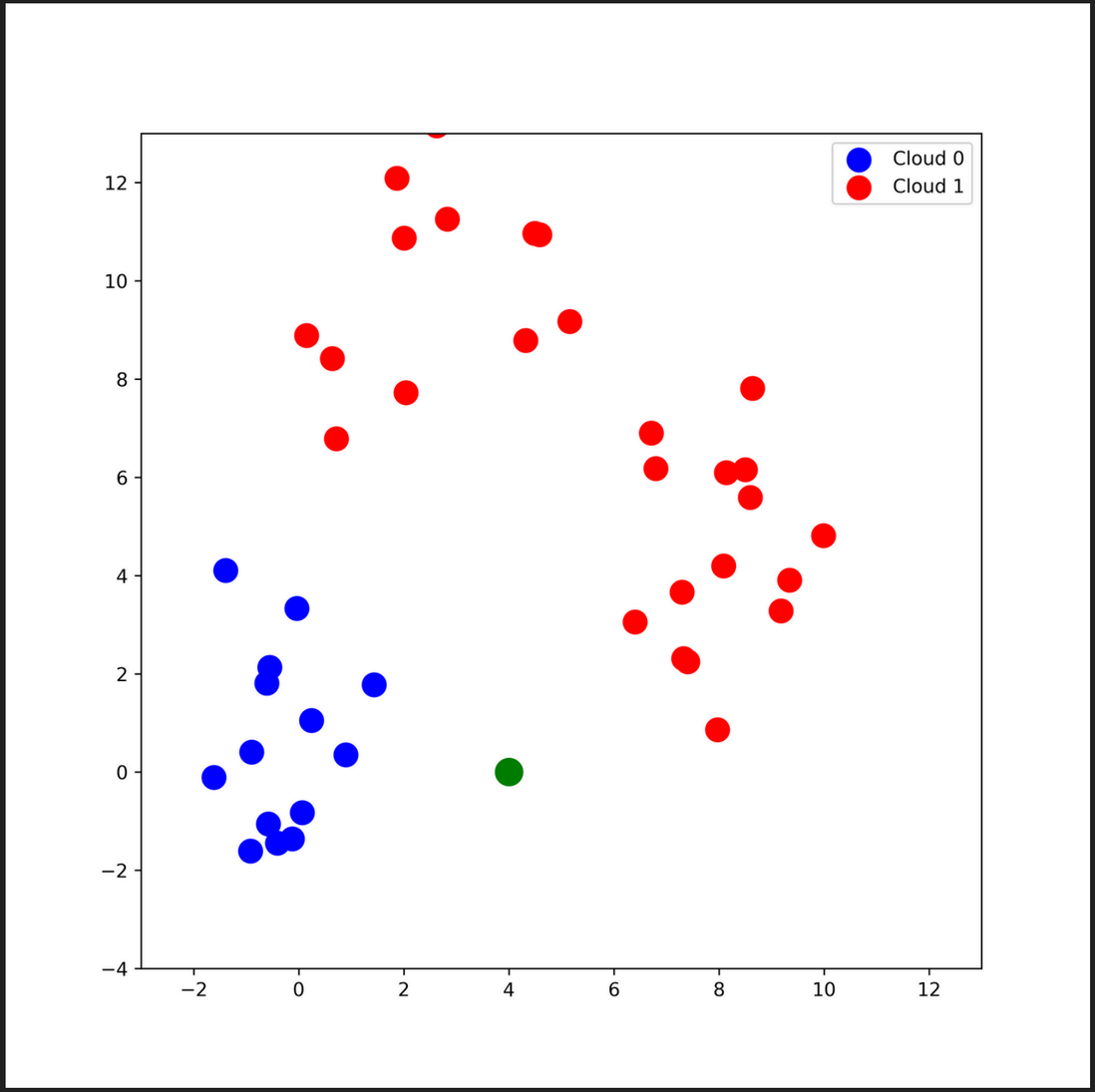
10.24432/C5V312

License

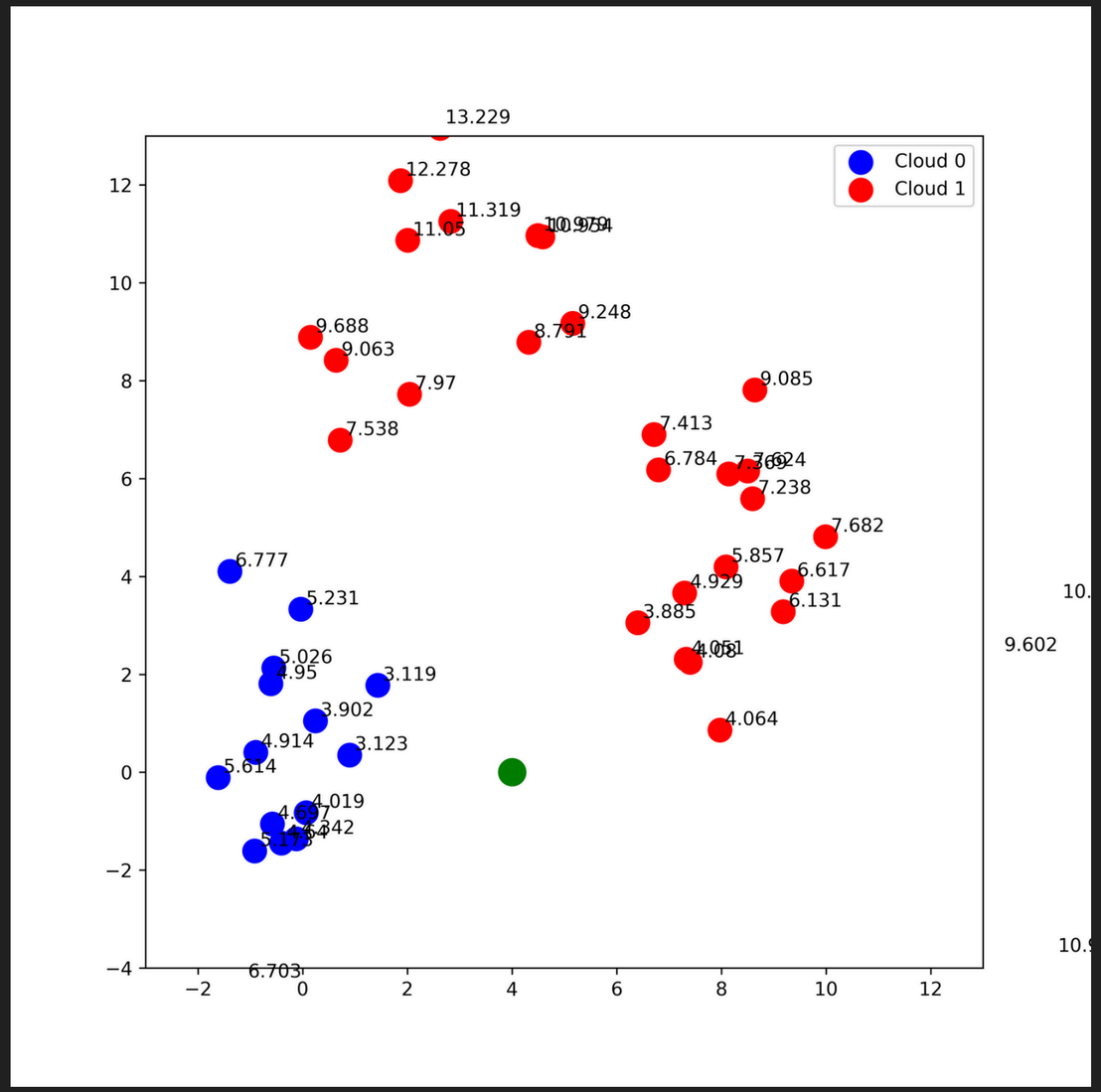
This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

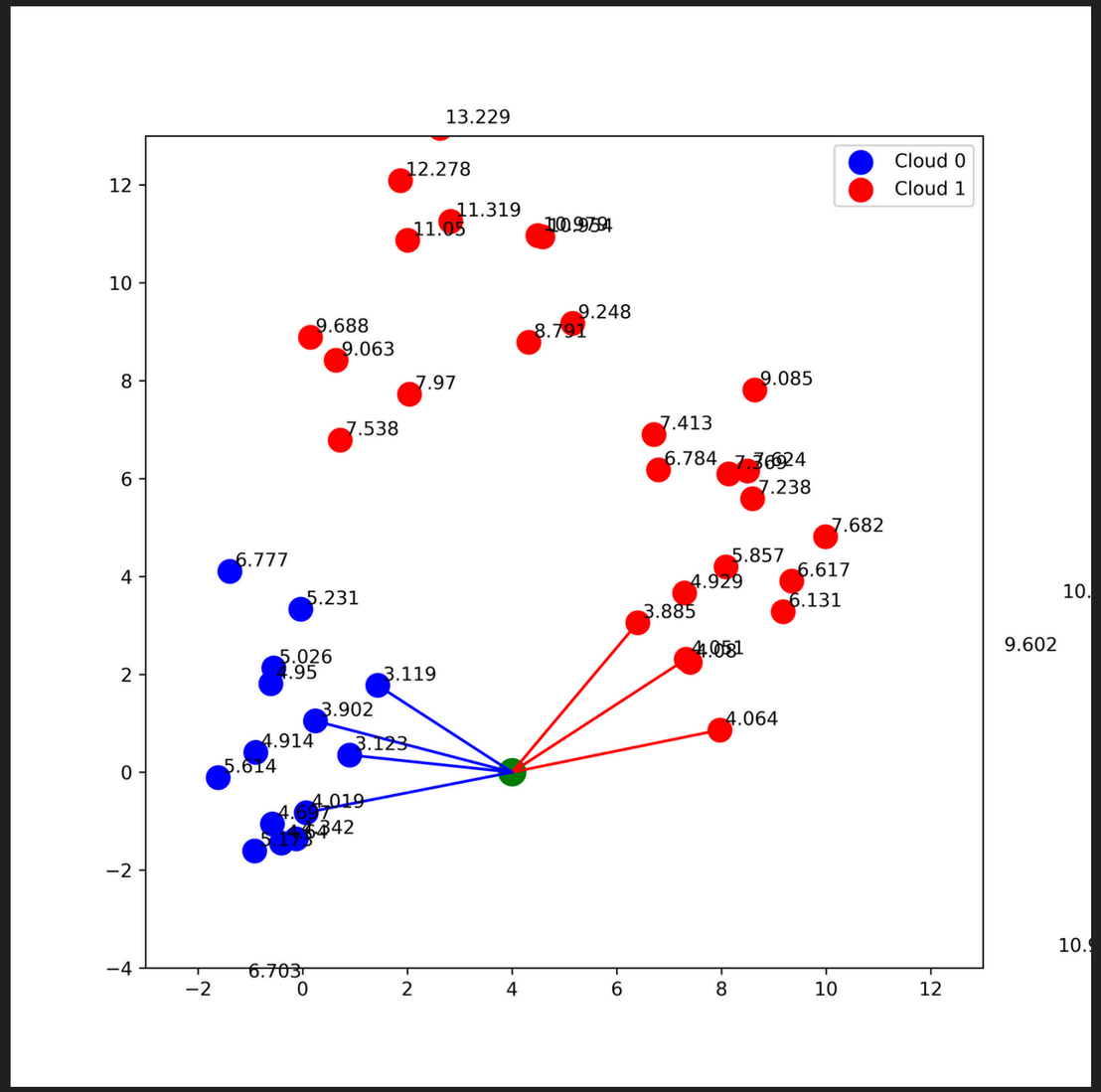
Algorytmy klasyfikacji: knn



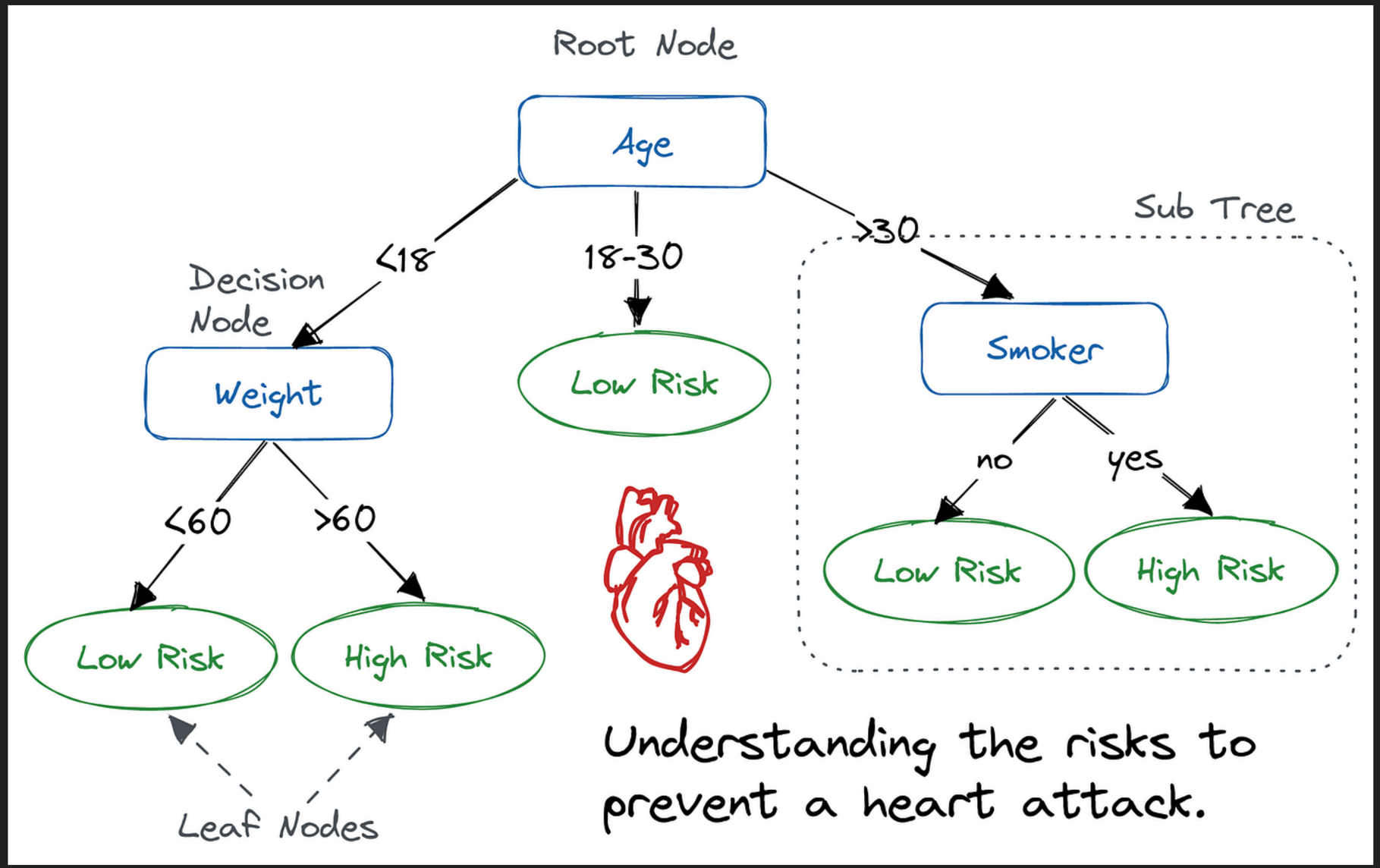
Algorytmy klasyfikacji: knn



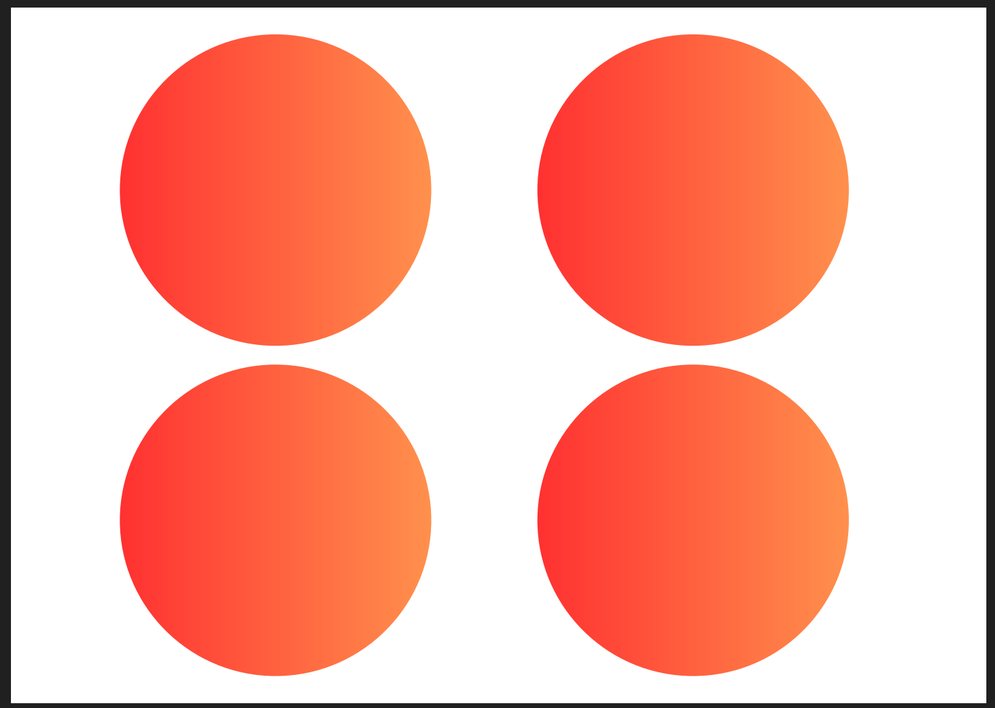
Algorytmy klasyfikacji: knn



Algorytmy klasyfikacji: drzewa decyzyjne

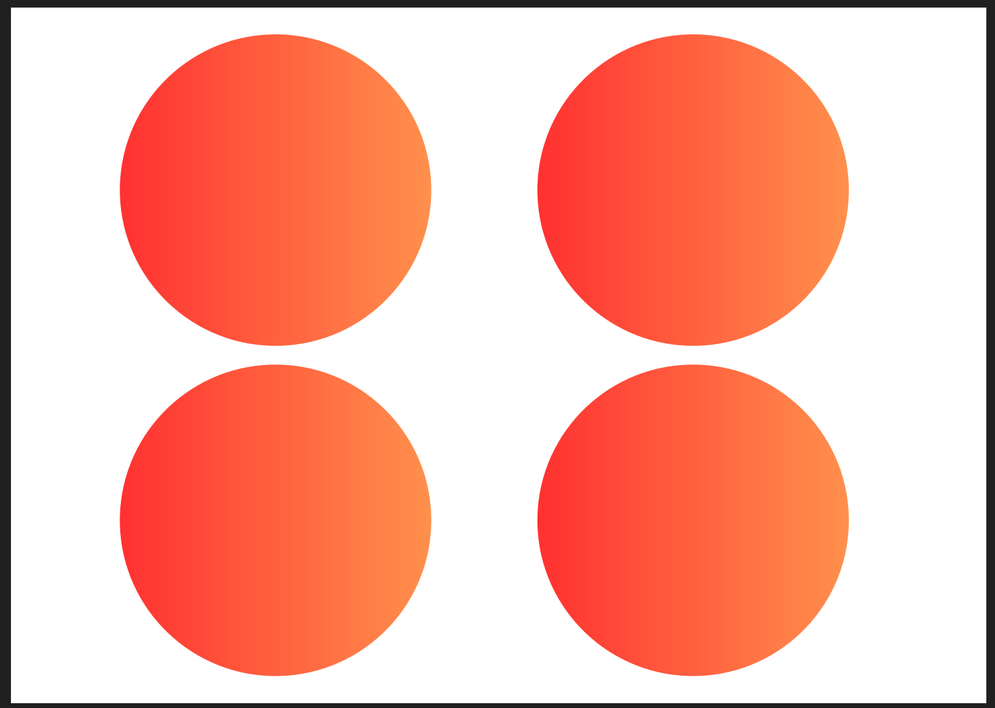


Algorytmy klasyfikacji: drzewa decyzyjne



Wysoka pewność
czyli
niska entropia

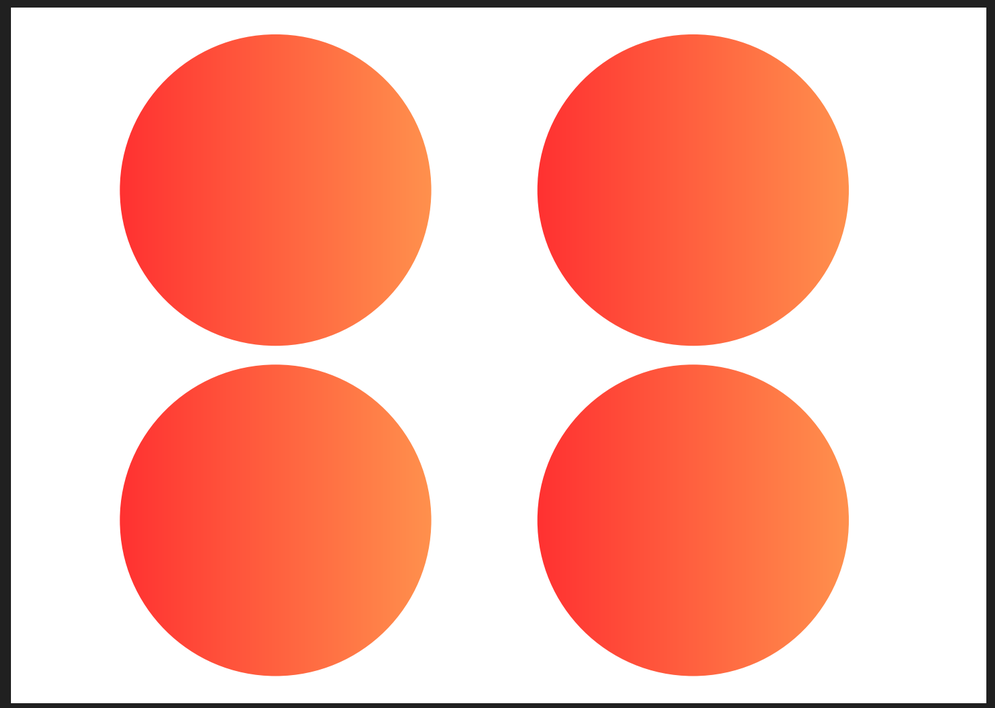
Algorytmy klasyfikacji: drzewa decyzyjne



Wysoka pewność
czyli
niska entropia

$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

Algorytmy klasyfikacji: drzewa decyzyjne

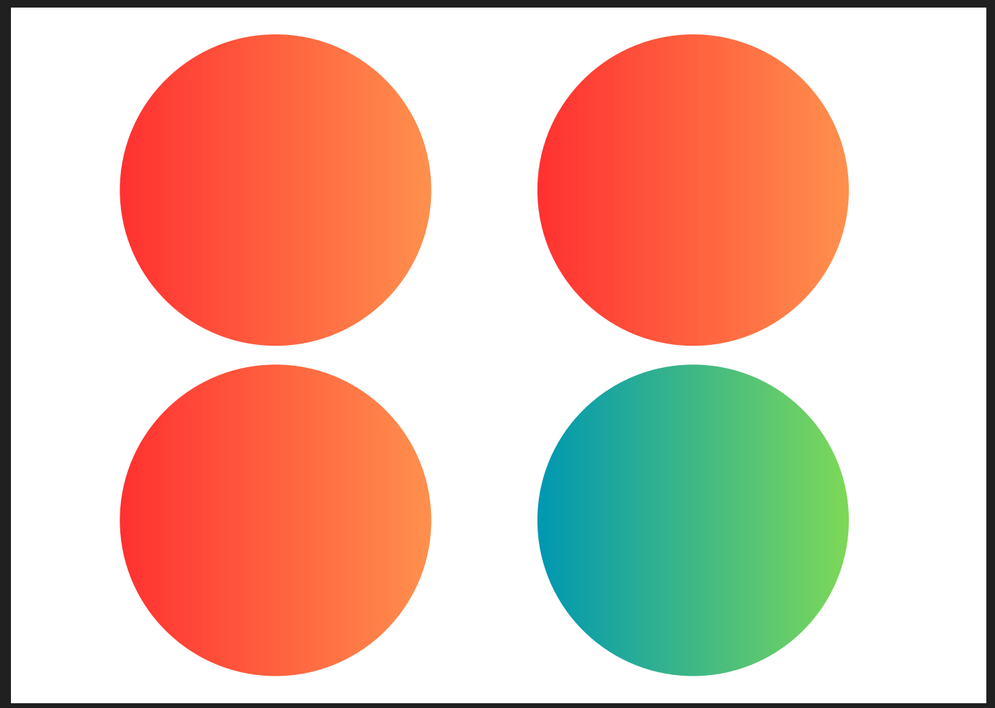


Wysoka pewność
czyli
niska entropia

$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

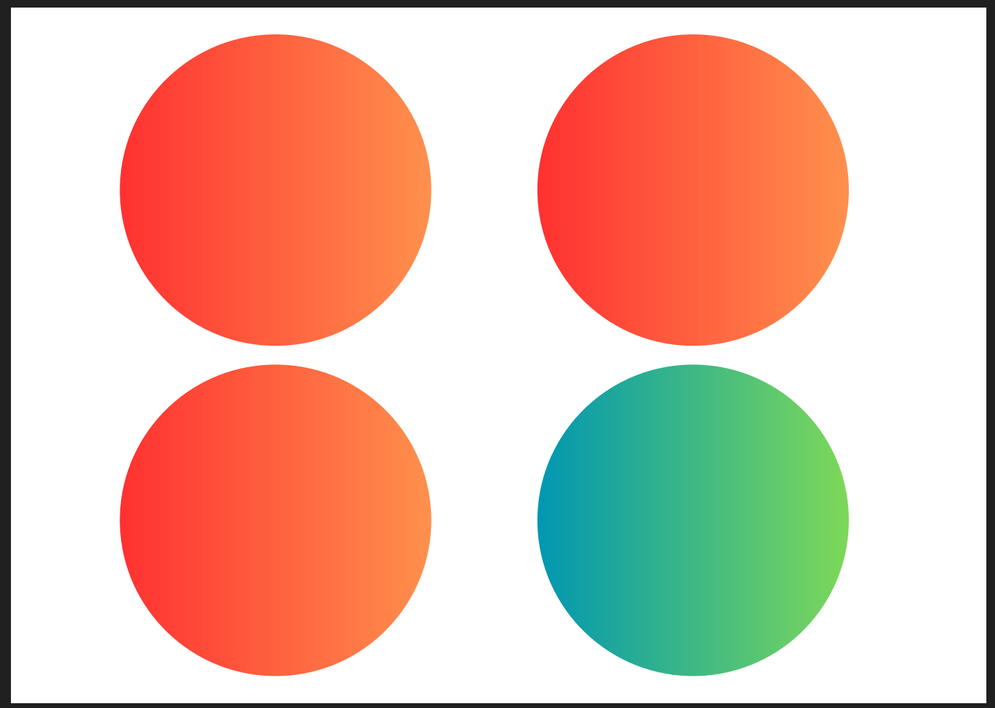
H = - (1*log2(1) + 0*log2(0)) = 0

Algorytmy klasyfikacji: drzewa decyzyjne



Średnia pewność
czyli
średnia entropia

Algorytmy klasyfikacji: drzewa decyzyjne

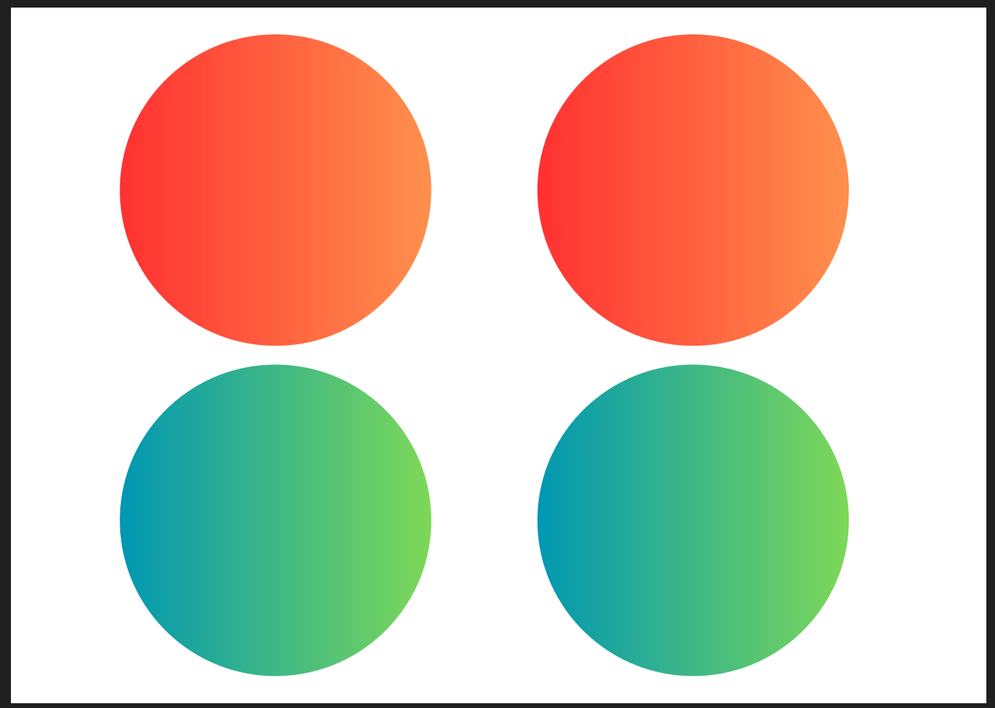


Średnia pewność
czyli
średnia entropia

$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

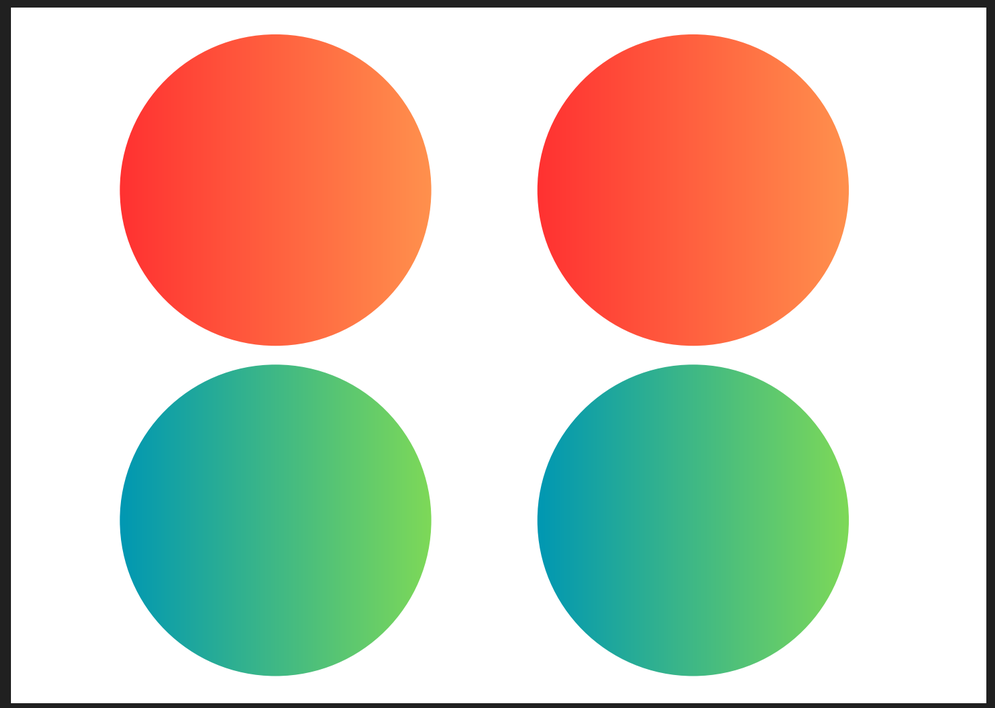
H = - (0.75*log2(0.75) + 0.25*log2(0.25)) = 0.81

Algorytmy klasyfikacji: drzewa decyzyjne



Niska pewność
czyli
wysoka entropia

Algorytmy klasyfikacji: drzewa decyzyjne

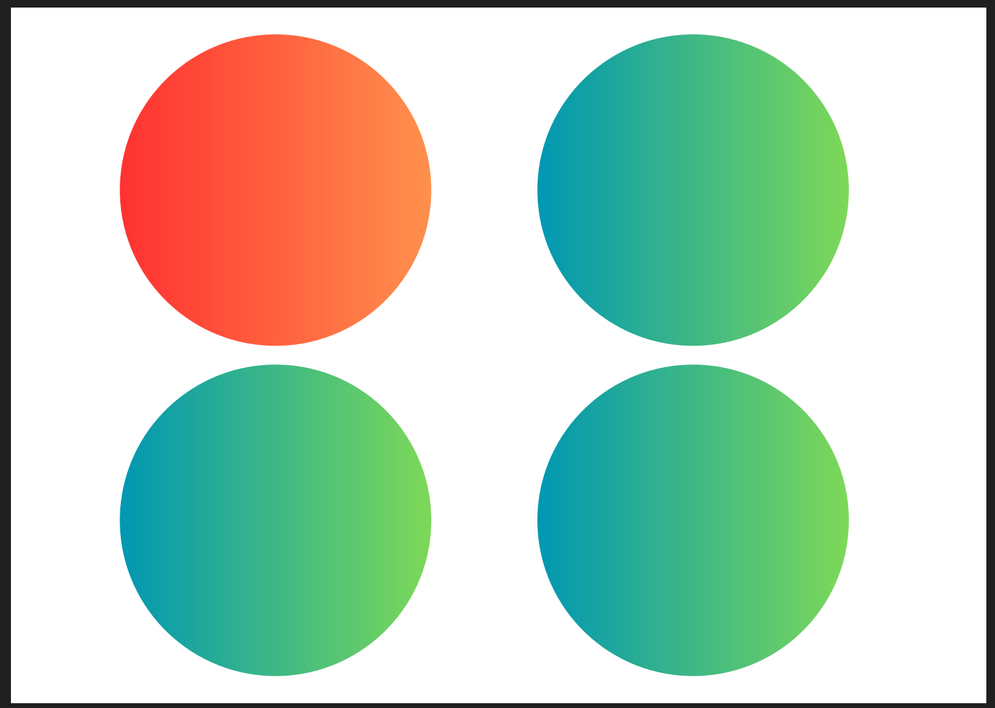


Niska pewność
czyli
wysoka entropia

$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

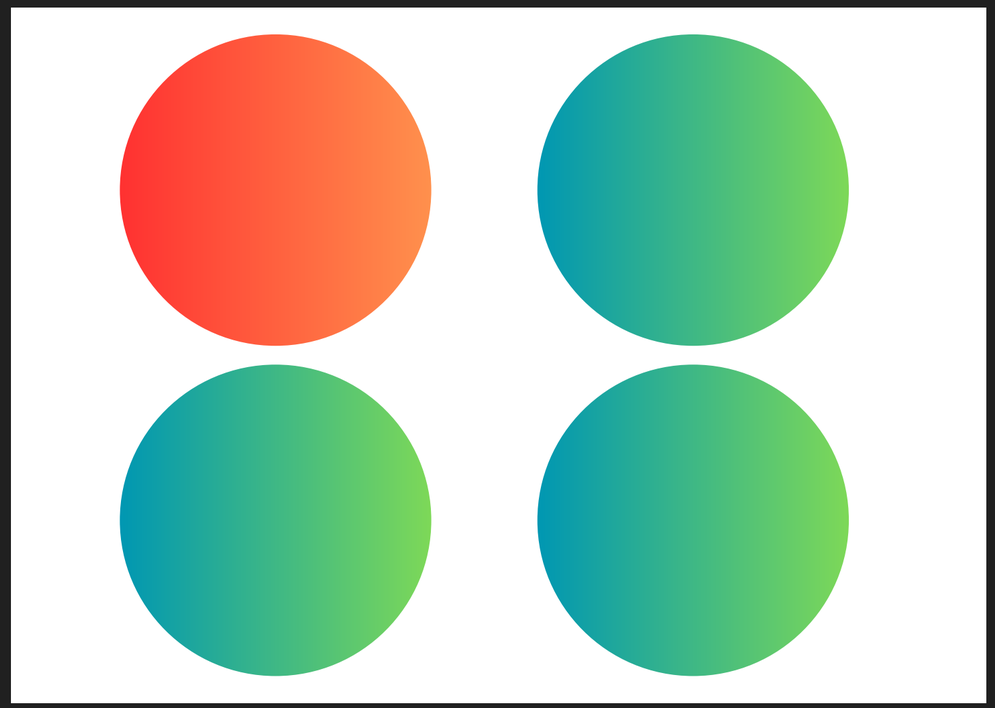
H = - (0.5*log2(0.5) +
0.5*log2(0.5)) = 1

Algorytmy klasyfikacji: drzewa decyzyjne



Średnia pewność
czyli
średnia entropia

Algorytmy klasyfikacji: drzewa decyzyjne

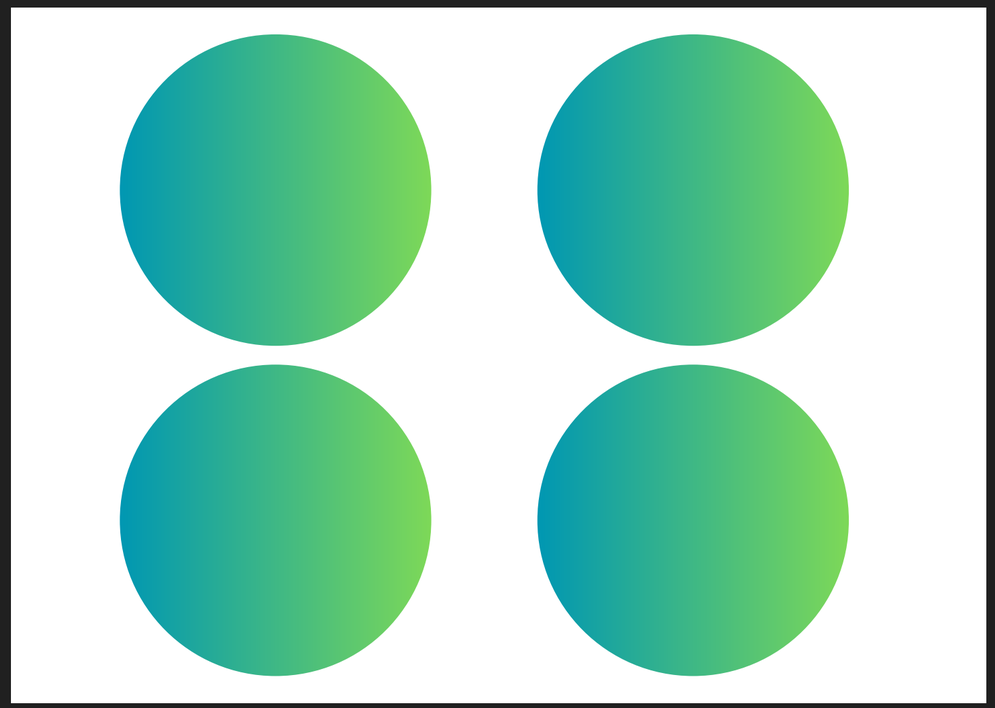


Średnia pewność
czyli
średnia entropia

$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

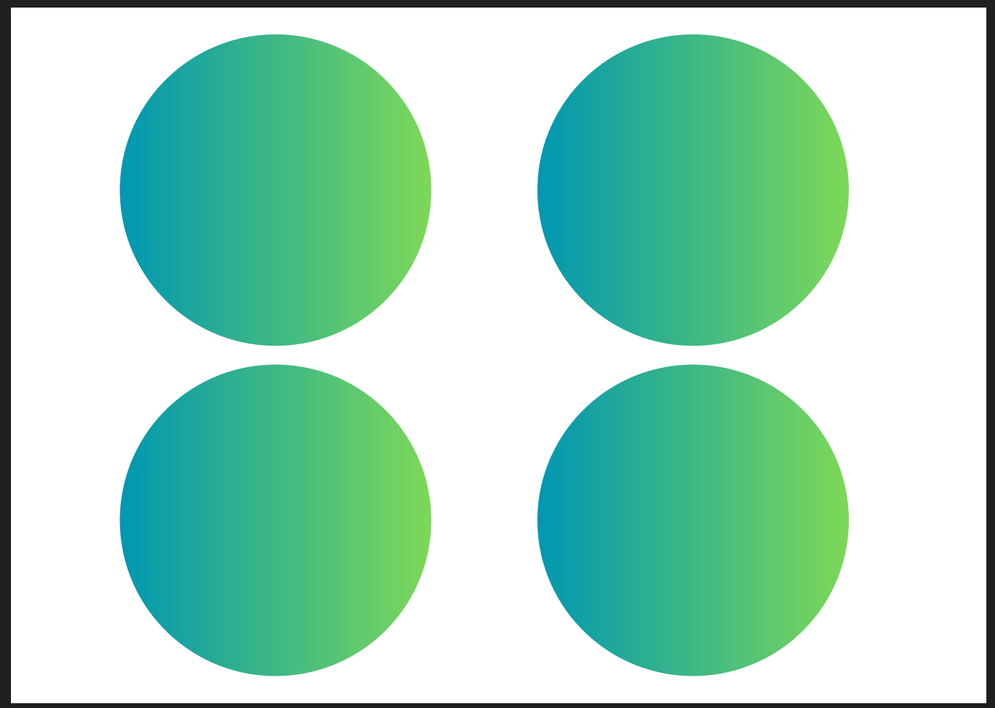
H = - (0.25*log2(0.25) +
0.75*log2(0.75)) = 0.81

Algorytmy klasyfikacji: drzewa decyzyjne



Wysoka pewność
czyli
niska entropia

Algorytmy klasyfikacji: drzewa decyzyjne

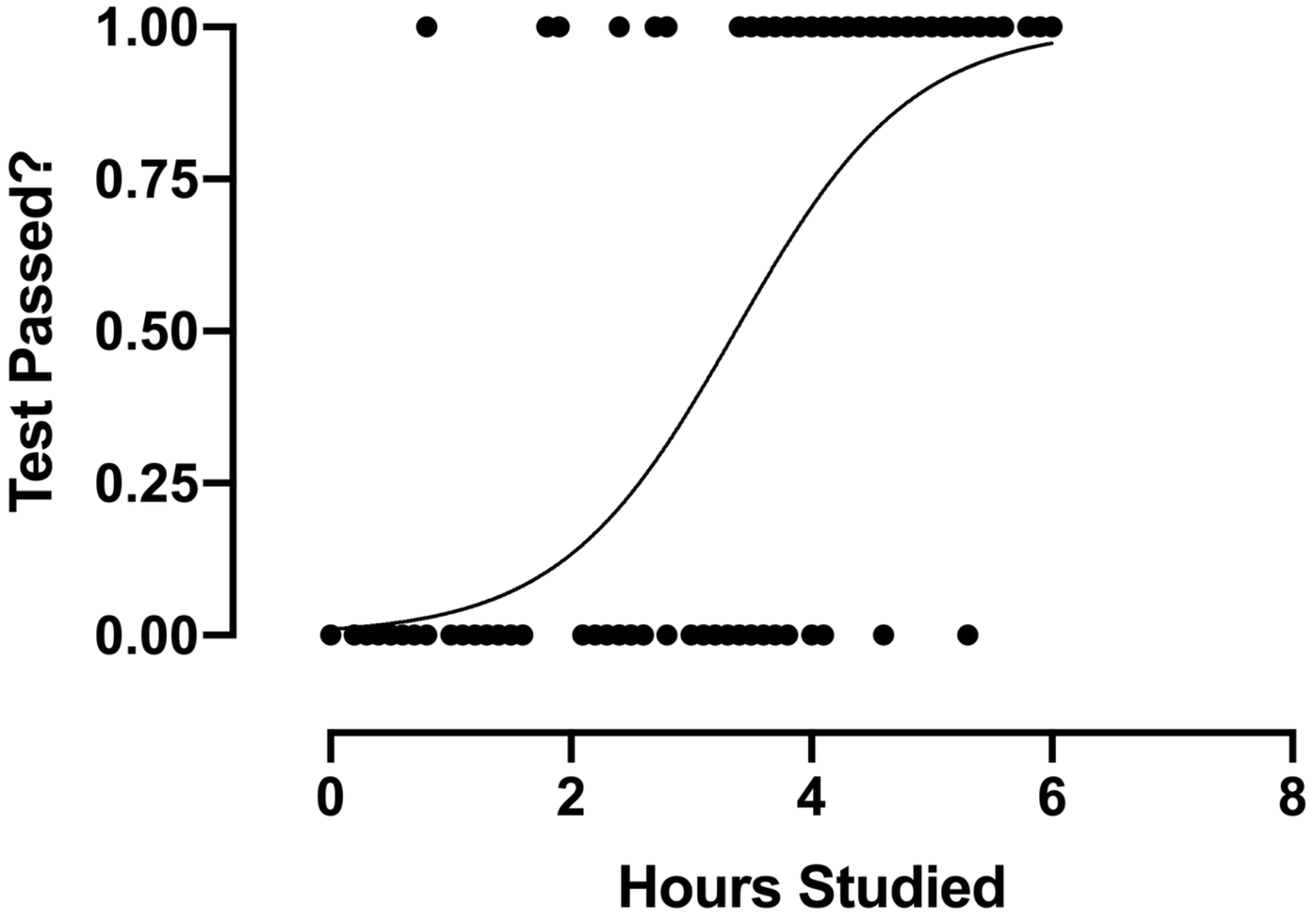


Wysoka pewność
czyli
niska entropia

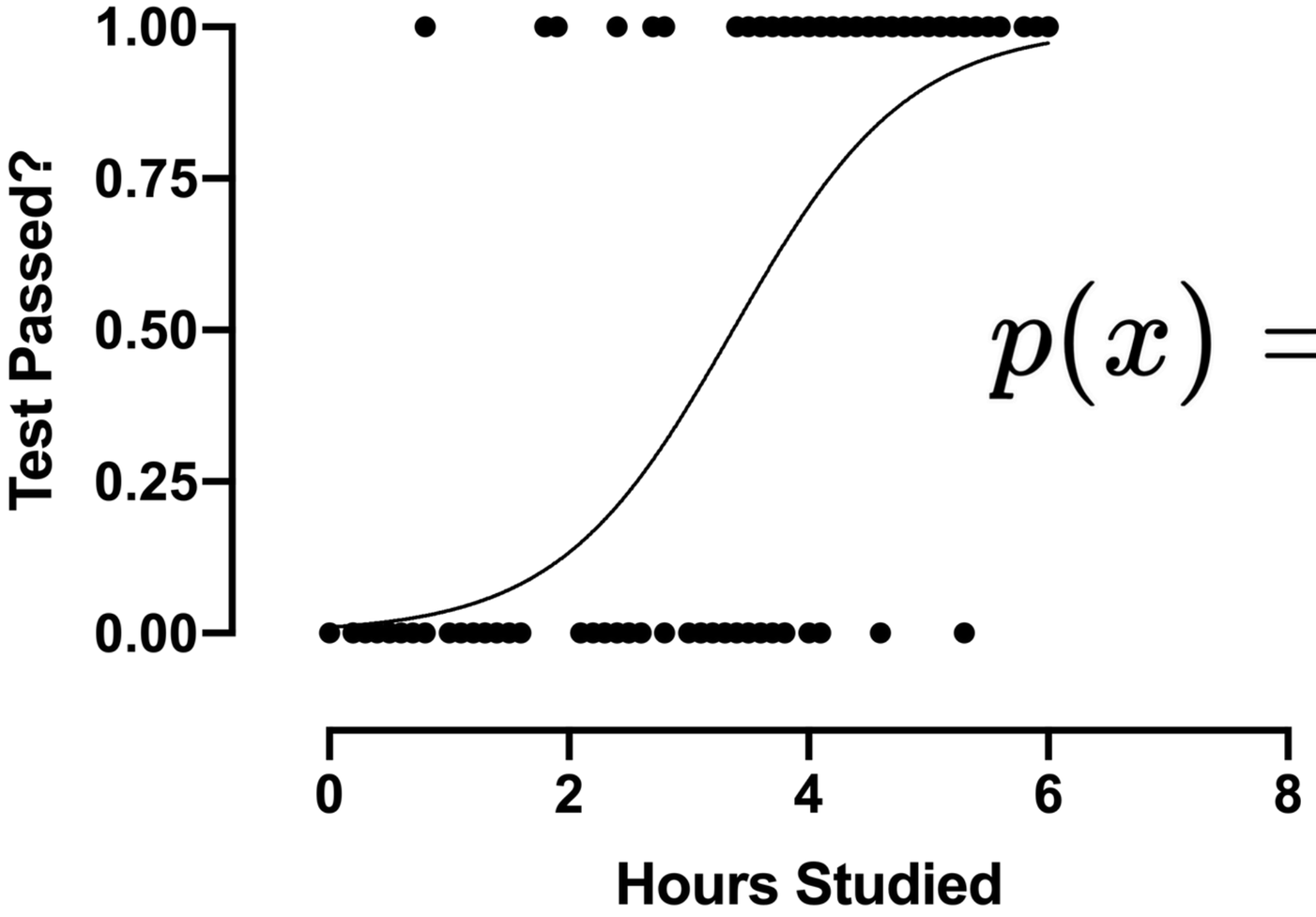
$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

H = - (1*log2(1) + 0*log2(0)) = 0

Algorytmy klasyfikacji: regresja logistyczna

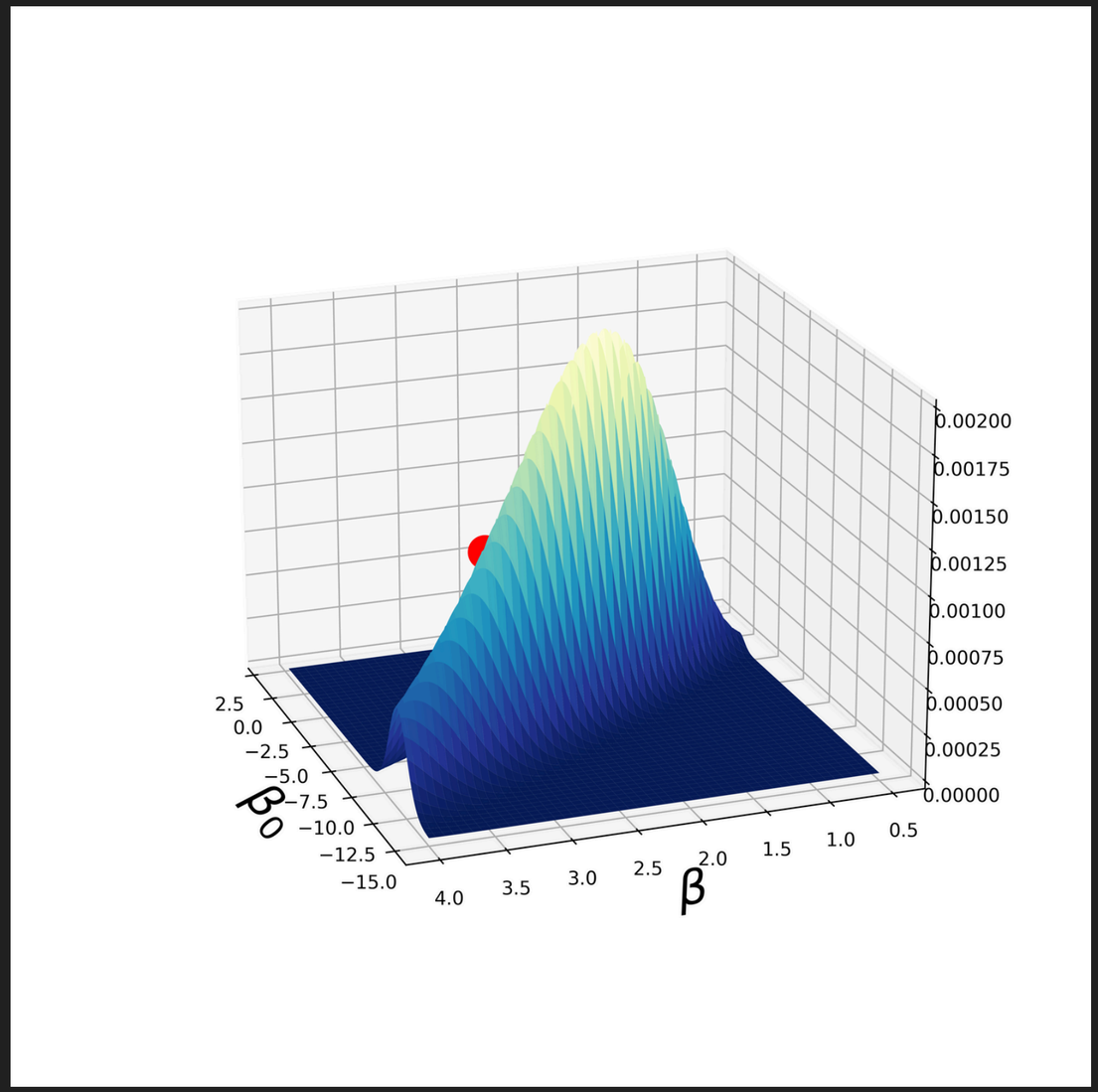


Algorytmy klasyfikacji: regresja logistyczna

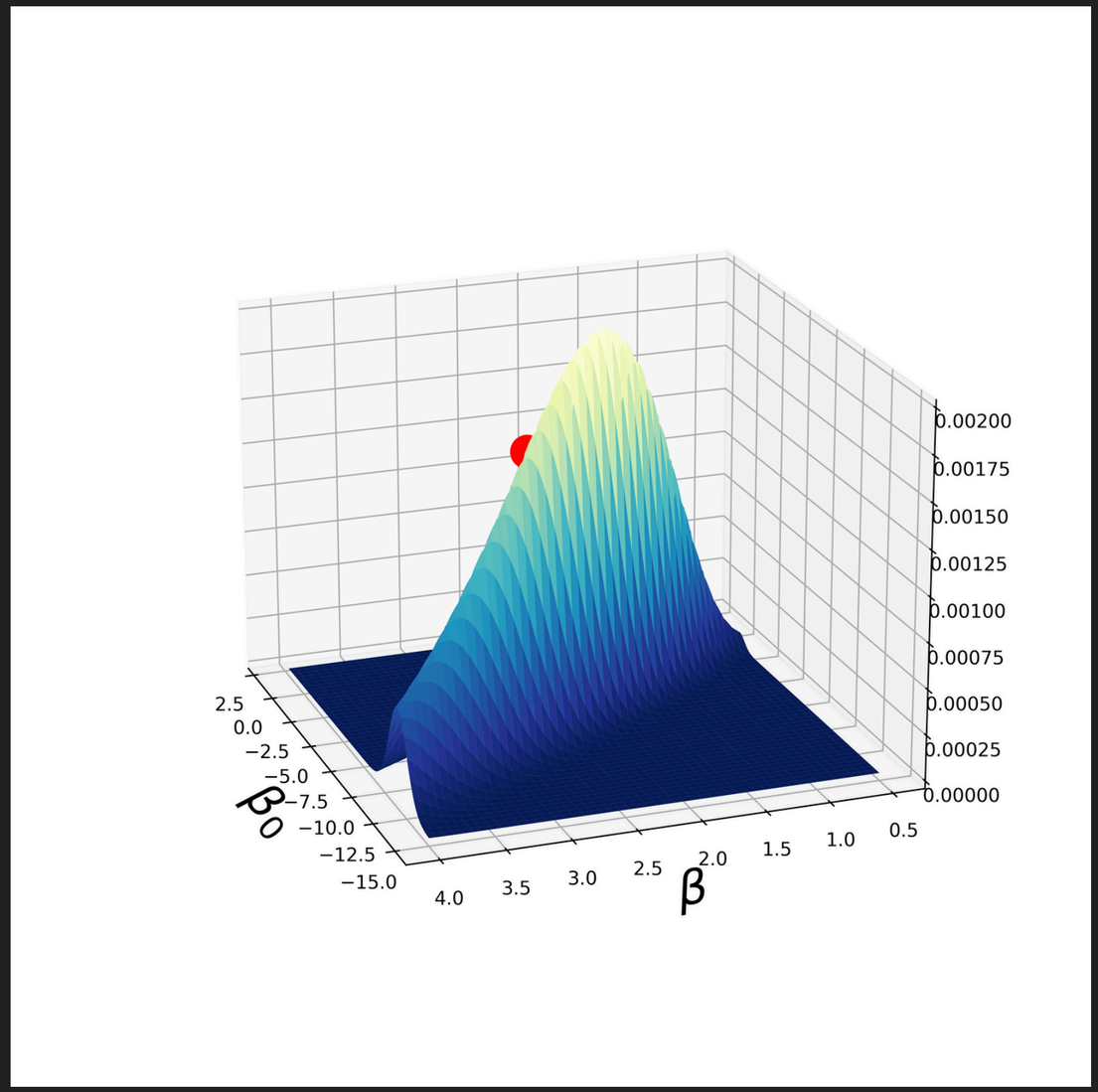


$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

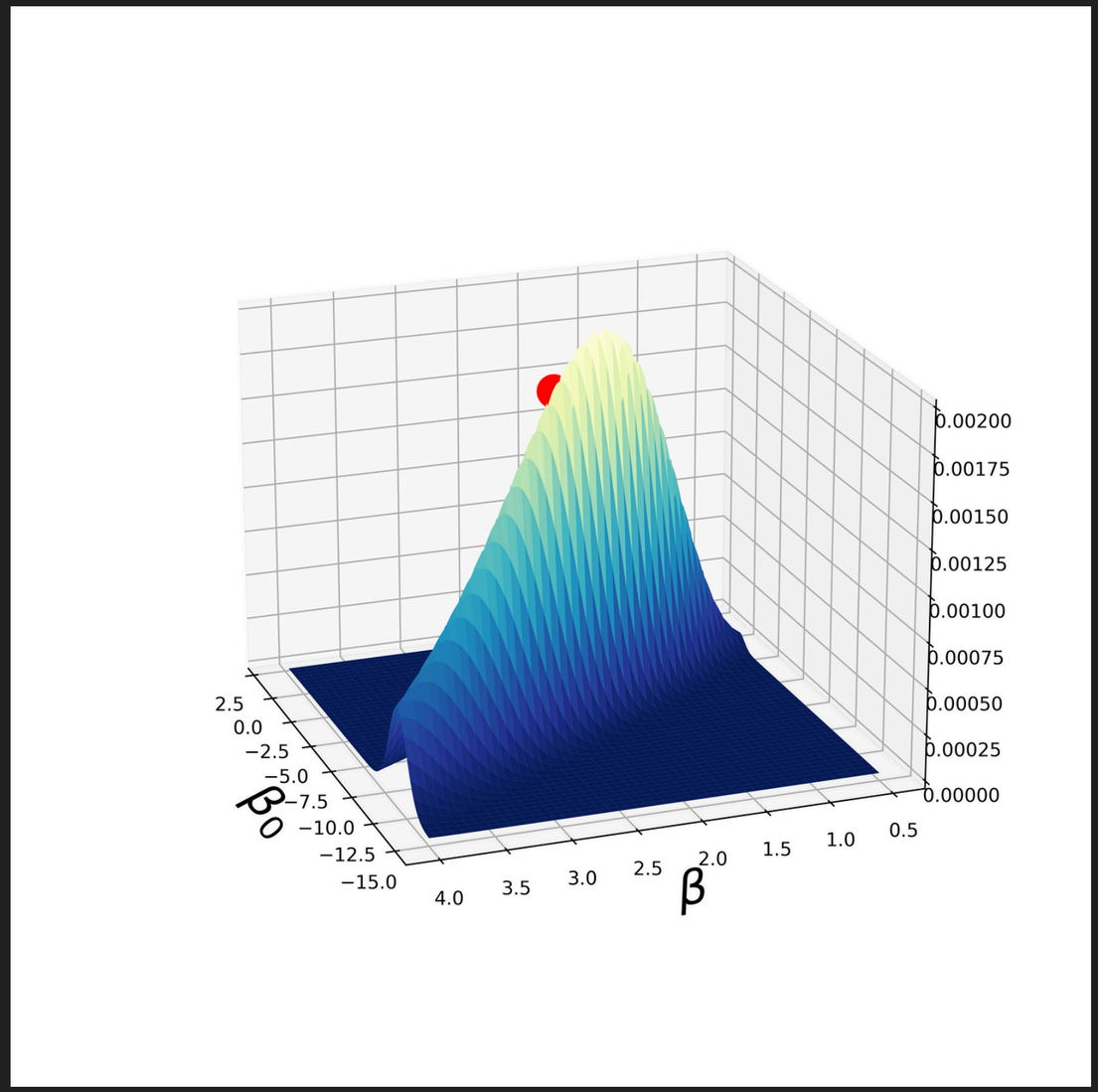
Algorytmy klasyfikacji: regresja logistyczna



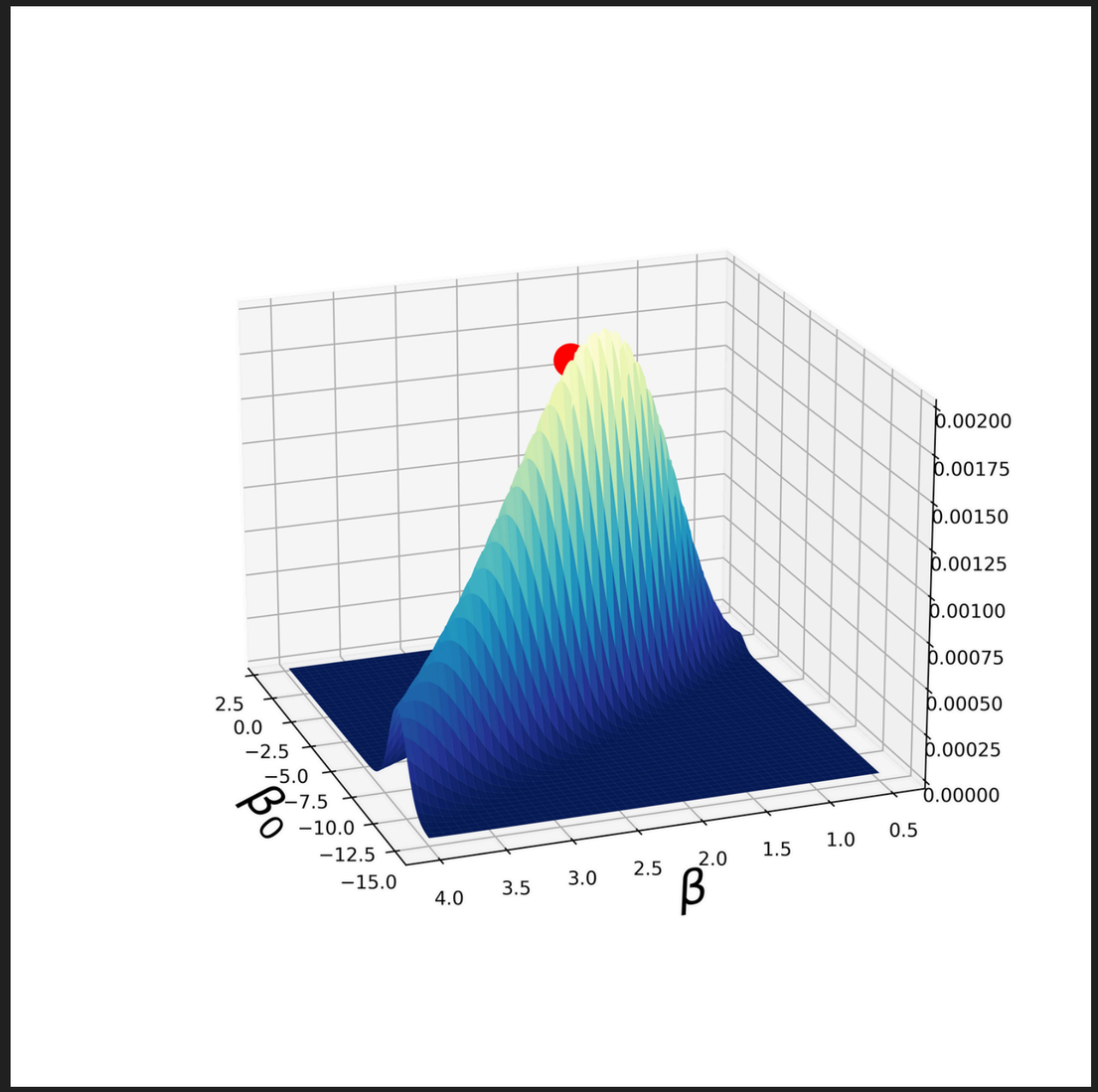
Algorytmy klasyfikacji: regresja logistyczna



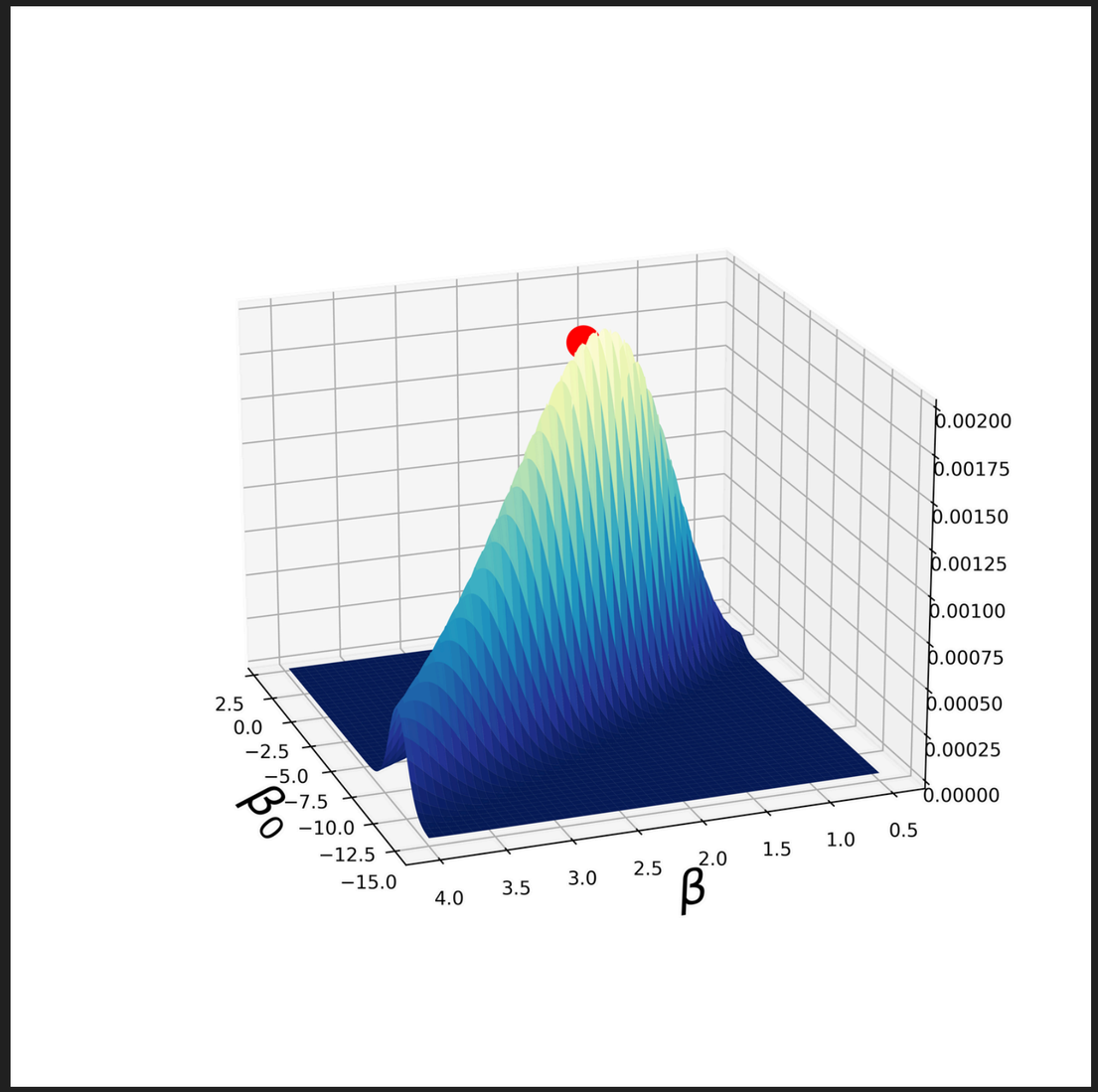
Algorytmy klasyfikacji: regresja logistyczna



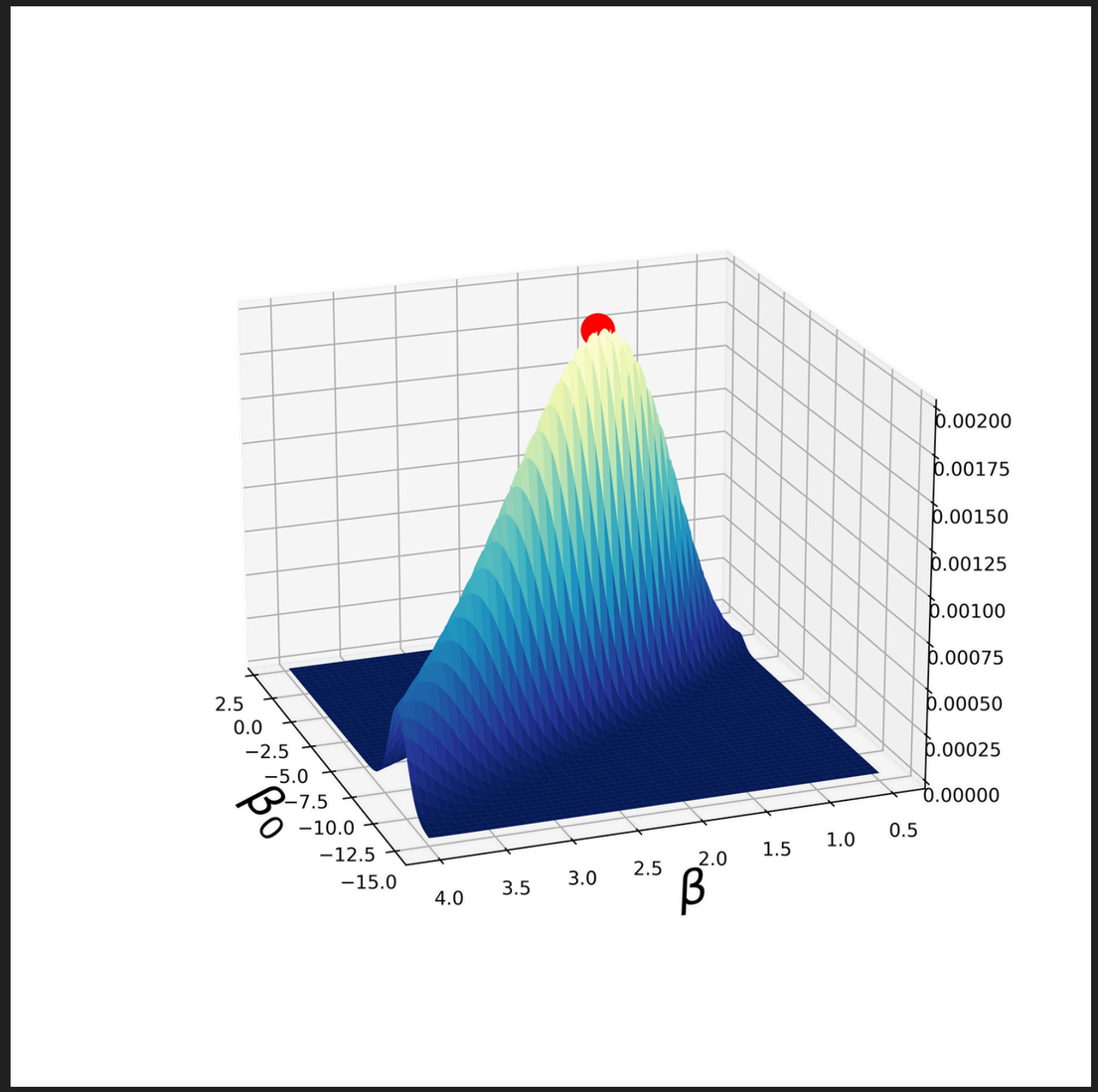
Algorytmy klasyfikacji: regresja logistyczna



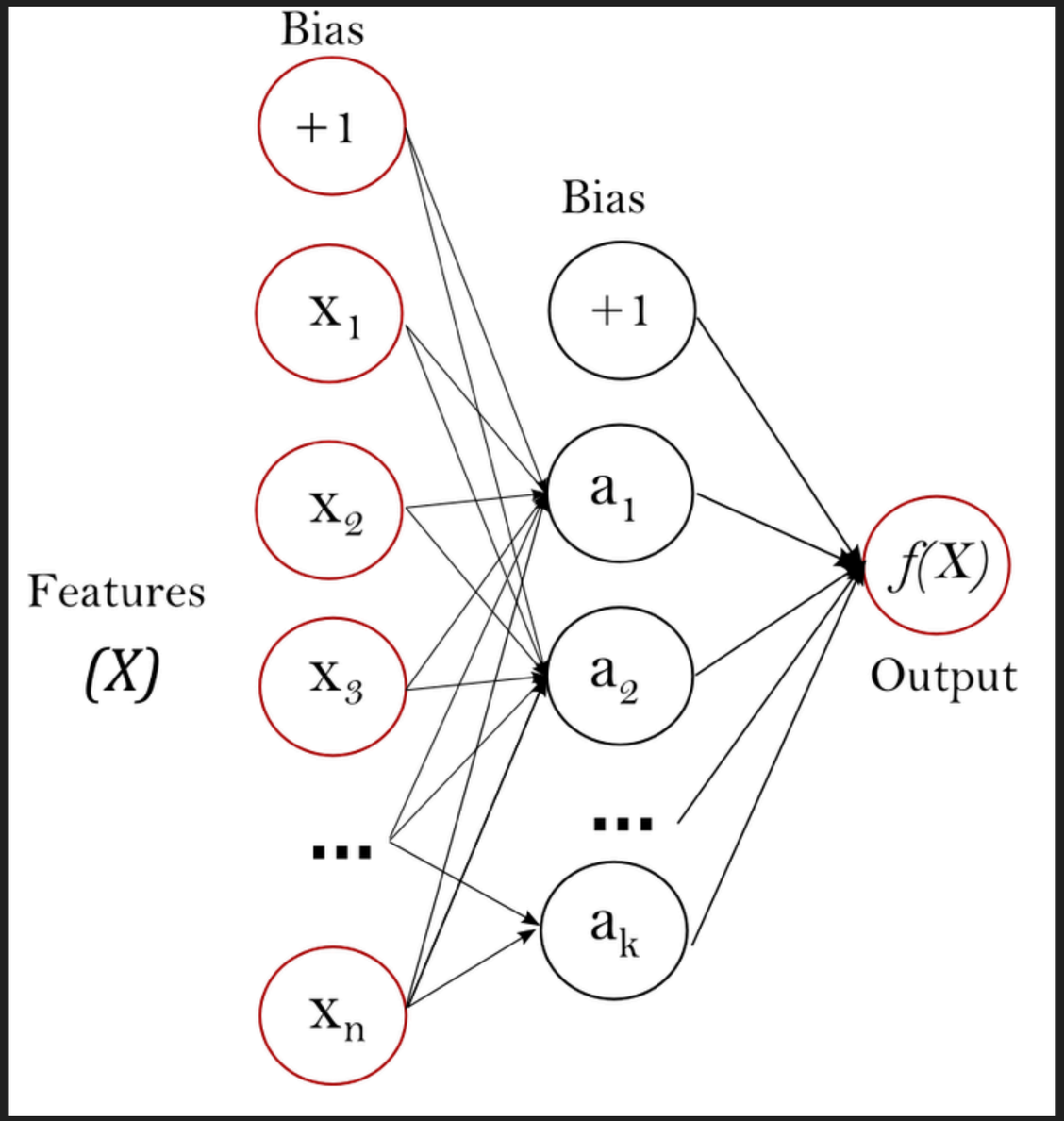
Algorytmy klasyfikacji: regresja logistyczna



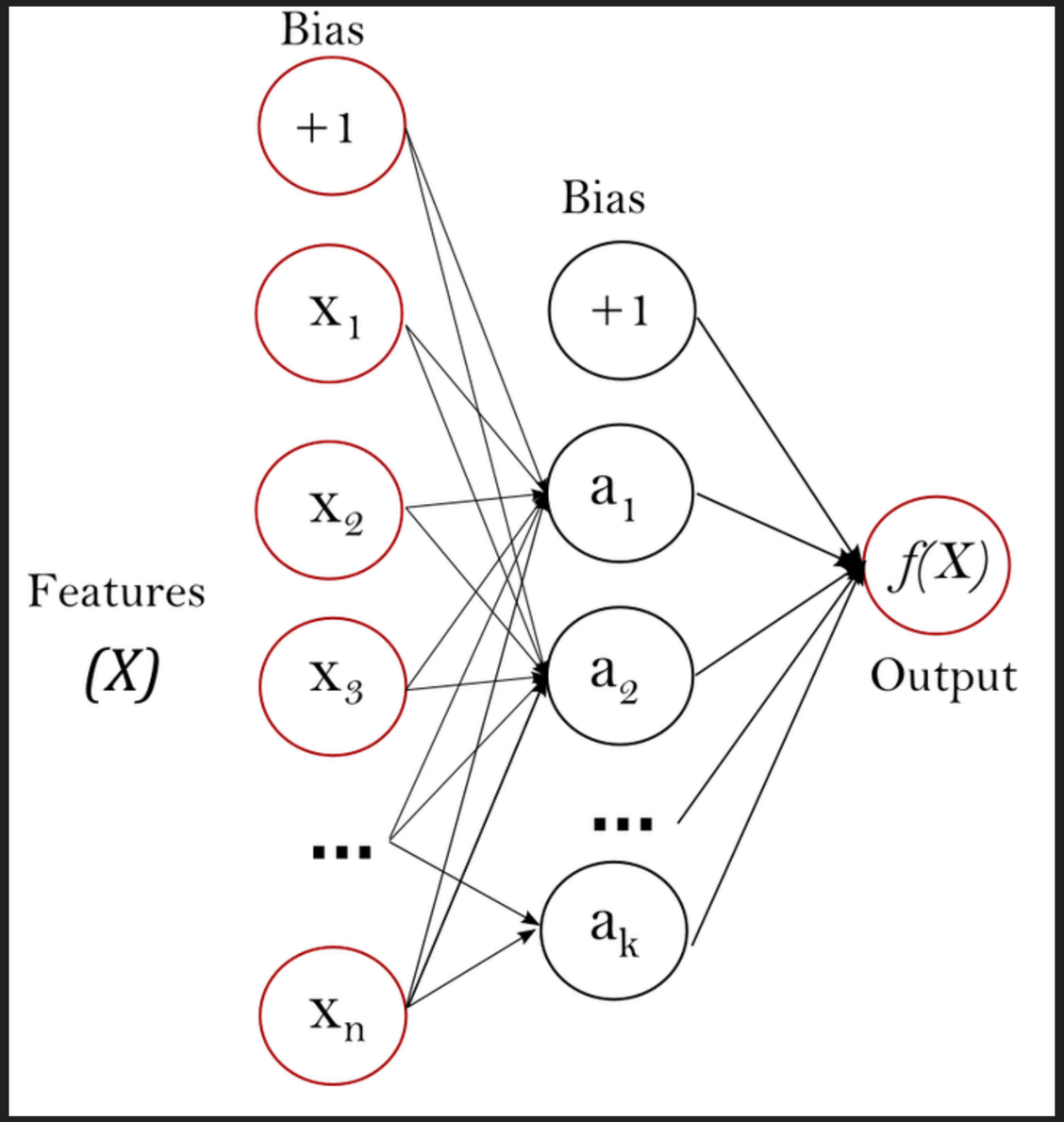
Algorytmy klasyfikacji: regresja logistyczna



Algorytmy klasyfikacji: sieci neuronowe



Algorytmy klasyfikacji: sieci neuronowe



Algorytmy klasyfikacji: sieci neuronowe