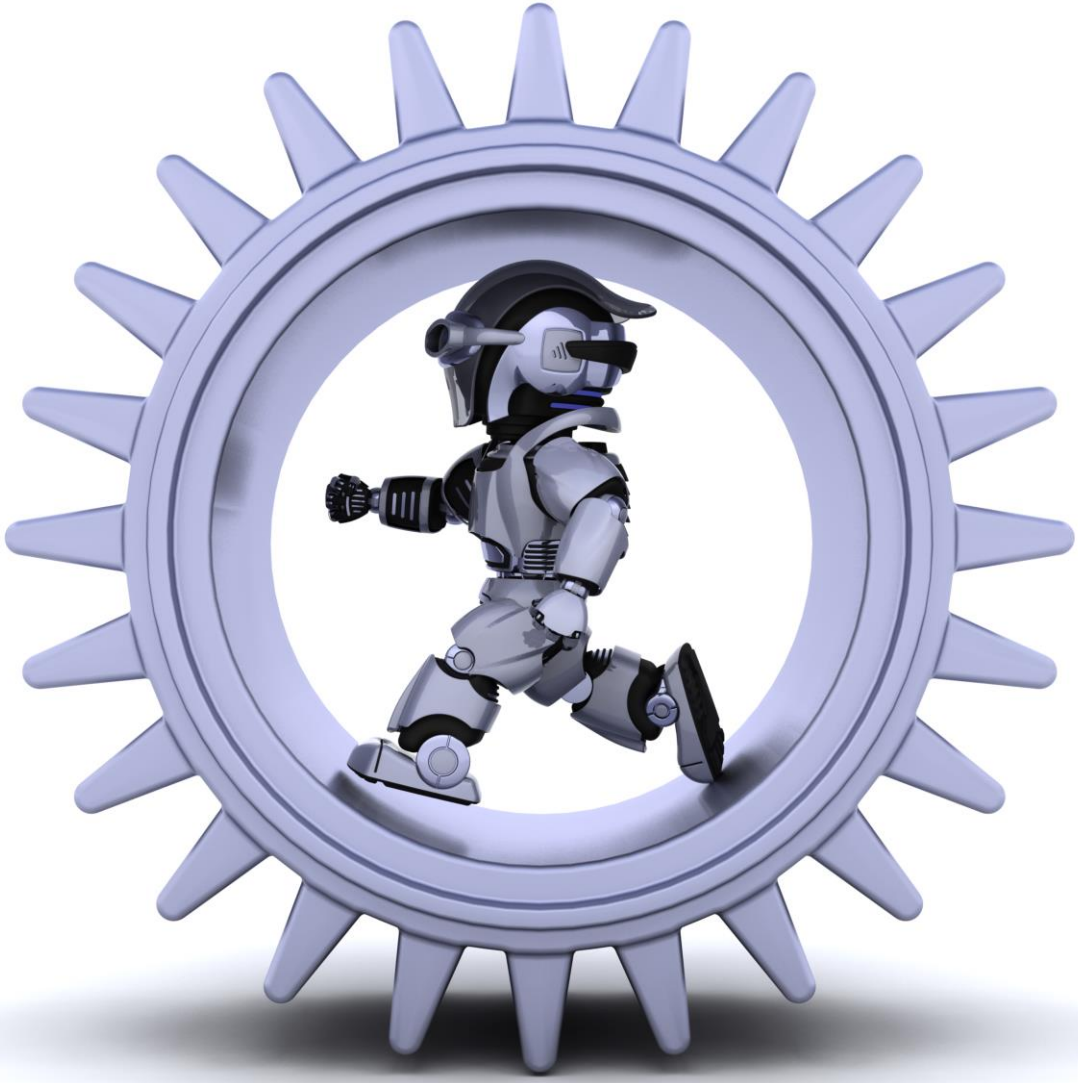
Podstawy uczenia maszynowego
PYTH_UM_T

Kamil Musiał

22-23.08.2024

Czym jest machine learning?



Machine Learning to zdolność komputerów do uczenia się bez programowania nowych umiejętności wprost.

Arthur Samuel (rok 1959)

Obraz autorstwa kjpargeter
freepik.com

Definicje machine learning

Definicja ogólna: Uczenie maszynowe to dziedzina sztucznej inteligencji, która polega na opracowywaniu algorytmów i modeli, które pozwalają komputerom na naukę na podstawie danych i dokonywanie przewidywań lub podejmowanie decyzji bez wyraźnego zaprogramowania dla konkretnego zadania.

Definicja matematyczna: Uczenie maszynowe to optymalizacja funkcji celu, gdzie dane wejściowe są używane do dopasowania modelu predykcyjnego, który minimalizuje różnicę między przewidywaniami a rzeczywistymi wynikami. Proces ten polega na modyfikacji parametrów modelu w celu minimalizacji błędu.

Definicje machine learning

Definicja z perspektywy inżynierii oprogramowania: Uczenie maszynowe to technika projektowania systemów komputerowych, które uczą się automatycznie na podstawie doświadczeń (danych) i poprawiają swoje działanie z czasem, bez potrzeby ręcznego programowania nowych reguł.

Definicja praktyczna: Uczenie maszynowe to narzędzie, które pozwala na tworzenie aplikacji, które mogą analizować dane, wyciągać wnioski i podejmować decyzje w sposób automatyczny. Jest ono używane w wielu dziedzinach, takich jak analiza obrazu, przetwarzanie języka naturalnego, i rekomendacje produktów.

Definicja z perspektywy biznesowej: Uczenie maszynowe to technologia, która umożliwia firmom przekształcanie danych w użyteczne informacje, które mogą być wykorzystane do podejmowania lepszych decyzji, automatyzacji procesów i poprawy efektywności operacyjnej.

ML vs AI

Zakres pojęciowy:

• **Sztuczna inteligencja (SI):** To szeroka dziedzina informatyki, której celem jest tworzenie systemów i algorytmów, które mogą naśladować lub symulować ludzką inteligencję. SI obejmuje różne techniki, metody i podejścia, które umożliwiają komputerom rozwiązywanie problemów, podejmowanie decyzji, przetwarzanie języka naturalnego, percepcję wizualną i inne zadania, które wymagają inteligencji.

• **Uczenie maszynowe (ML):** To podzbiór sztucznej inteligencji, skoncentrowany na tworzeniu i stosowaniu algorytmów, które umożliwiają systemom uczenie się na podstawie danych. ML skupia się na budowaniu modeli, które poprawiają swoje działanie wraz z dostępem do większej ilości danych, bez konieczności ręcznego programowania konkretnych reguł.

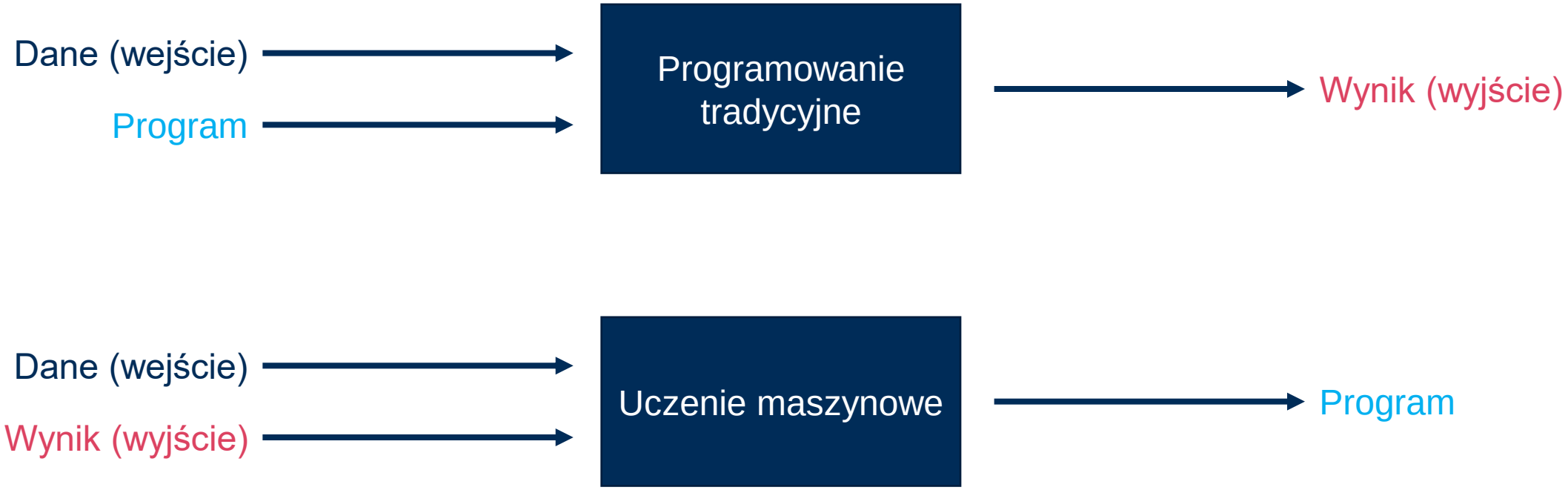
ML vs AI

Metodologia:

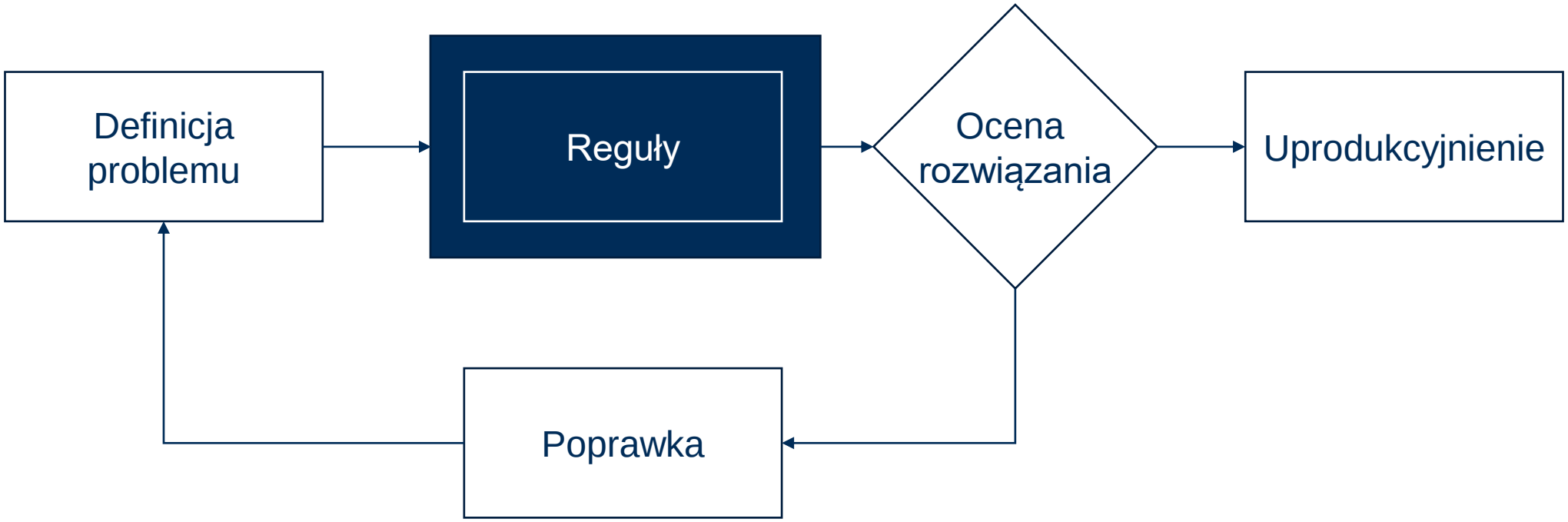
• **Sztuczna inteligencja (SI):** Obejmuje różne podejścia, takie jak logika rozmyta, systemy ekspertowe, algorytmy ewolucyjne, przetwarzanie języka naturalnego, systemy rekomendacyjne, a także uczenie maszynowe. Jest to więc dziedzina, która używa szerokiego wachlarza technik, aby symulować inteligencję.

• **Uczenie maszynowe (ML):** Skupia się na technikach statystycznych i matematycznych do analizy danych. W ML używa się modeli takich jak regresja liniowa, drzewa decyzyjne, sieci neuronowe, a także technik takich jak uczenie nadzorowane, nienadzorowane i wzmacniające.

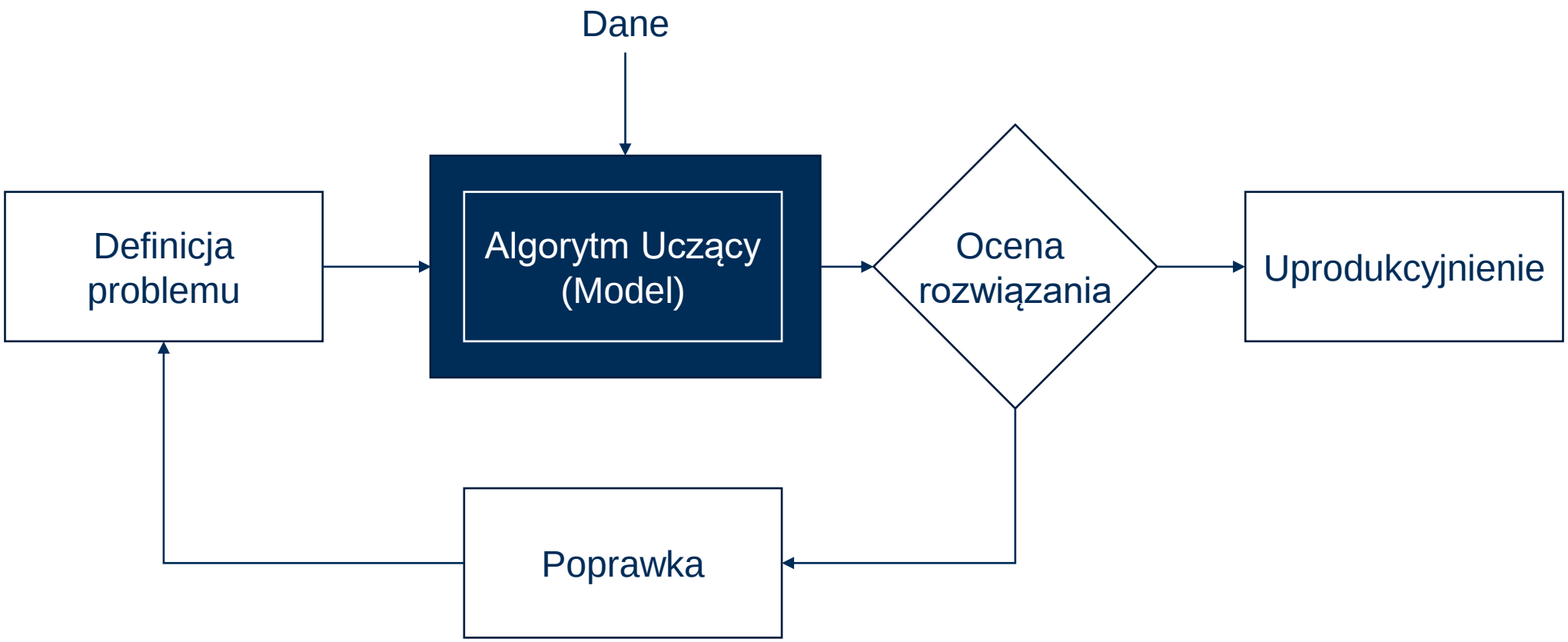
Jak działa machine learning?



Jak działa machine learning?



Jak działa machine learning?



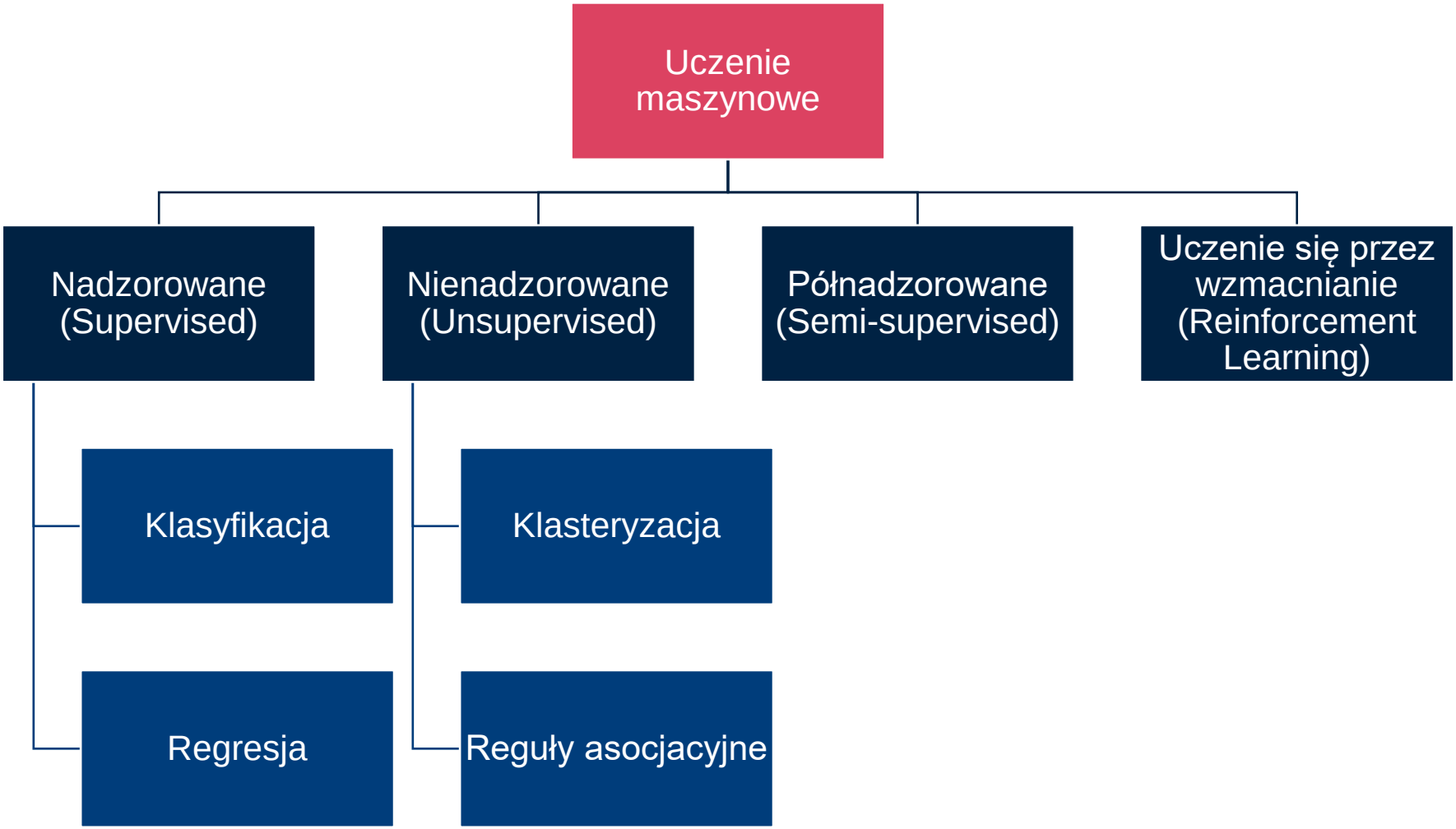
Proces machine learning?

- Gromadzenie danych z przeszłości w dowolnej formie odpowiedniej do przetwarzania.
  Im lepsza jakość danych, tym bardziej nadaje się do modelowania
- Przetwarzanie danych – etykietowanie, usuwanie wartości pustych, wypełnianie pustych wartości, normalizacja, poszukiwanie korelacji
- Wybór modelu
- Wygenerowanie danych testowych
- Trenowanie modelu
- Weryfikacja działania

Zastosowanie machine learning?

- Predykcja zdarzeń/cen/wartości
- Kategoryzacja tekstów, obrazów, badań
- Rozpoznawanie nowych obiektów
- Analiza zachowań użytkowników (rekomendacje)
- Wykrywanie anomalii (np. fraudów)

Rodzeje machine learning?



Klasyfikacja uczenia maszynowego

- **Uczenie nadzorowane**

Algorytm uczy się na podstawie danych przykładowych i powiązanych odpowiedzi docelowych, które mogą składać się z wartości liczbowych lub etykiet ciągów, takich jak klasy lub znaczniki, w celu późniejszego przewidzenia poprawnej odpowiedzi, gdy zostaną przedstawione nowe przykłady, należy do kategorii Uczenie nadzorowane. Takie podejście jest rzeczywiście podobne do uczenia się przez człowieka pod nadzorem nauczyciela. Nauczyciel podaje dobre przykłady do zapamiętania przez ucznia, a następnie uczeń czerpie ogólne zasady z tych konkretnych przykładów.

- **Uczenie bez nadzoru**

Algorytm uczy się na podstawie prostych przykładów bez żadnej powiązanej odpowiedzi, pozostawiając algorytmowi samodzielne określenie wzorców danych. Ten typ algorytmu ma tendencję do przekształcania danych w coś innego, na przykład nowe funkcje, które mogą reprezentować klasę lub nową serię nieskorelowanych wartości. Są one bardzo przydatne w zapewnianiu wglądu w znaczenie danych i nowych przydatnych danych wejściowych do nadzorowanych algorytmów uczenia maszynowego.

Klasyfikacja uczenia maszynowego

- **Uczenie ze wzmocnieniem**

Uczenie ze wzmocnieniem to trzeci popularny typ uczenia maszynowego. Wykorzystuje obserwacje zebrane w wyniku interakcji z otoczeniem do podjęcia działań, które zmaksymalizują nagrodę lub zminimalizują ryzyko. W tym przypadku algorytm uczenia ze wzmocnieniem (zwany agentem) stale uczy się ze swojego otoczenia za pomocą iteracji. Doskonałym przykładem wzmacniania uczenia są komputery osiągające nadzwyczajną biegłość pozwalającą im pokonywać ludzi w grach komputerowych.

Regresja liniowa

- **Regresja**

Algorytm pierwszego wyboru przy prognozowaniu wartości ciągłych. Jeden z najpopularniejszych i najprostszych algorytmów. Zakłada on istnienie zależności liniowej pomiędzy zmienną modelowaną a predyktorem/-ami. W najprostszym przypadku regresji liniowej przedstawia ona jedną zmienną modelowaną i jeden predyktor. Zależność pomiędzy nimi jest modelowana na dwuwymiarowym układzie współrzędnych za pomocą prostej o odpowiednim nachyleniu. Odzwierciedla ona trend i zmienność danych. Rozwiązaniem problemu prognozowania z użyciem regresji liniowej będą odpowiednio dobrane parametry.

- **Regresja liniowa**

Regresja liniowa jest najprostszym wariantem regresji w statystyce. Zakłada ona, że zależność pomiędzy zmienną objaśnianą a objaśniająca jest zależnością liniową. W regresji liniowej zakłada się, że wzrostowi jednej zmiennej (predyktor/predyktory) towarzyszy wzrost lub spadek na drugiej zmiennej. Analiza regresji liniowej ma na celu wyliczenie takich współczynników regresji (współczynników w modelu liniowym), aby model jak najlepiej przewidywał wartość zmiennej zależnej – tzn. błąd oszacowania był jak najmniejszy.

Regresja logistyczna

- **Regresja logistyczna**

Regresja logistyczna jest jednym z najprostszych i najczęściej stosowanych algorytmów uczenia maszynowego dla klasyfikacji dwóch klas. Jest łatwa w implementacji i może być użyta jako punkt odniesienia dla dowolnego problemu klasyfikacji binarnej. Regresja logistyczna opisuje i szacuje związek między jedną zależną zmienną binarną i zmiennymi niezależnymi.

- Losowy dobór próby
- Uwzględnienie wszystkich istotnych zmiennych
- Wyłączenie z modelu wszystkich nieistotnych zmiennych
- Zmienne niezależne nie mogą być współliniowe
- Regresja logistyczna jest wrażliwa na występowanie punktów odstających
- Próba musi być dostatecznie liczna (co najmniej n=100)

Rodzaje algorytmów ML

- **Klasyfikacja**

Dane wejściowe są podzielone na dwie lub więcej klas, a uczeń musi stworzyć model, który przypisuje niewidoczne dane wejściowe do jednej lub więcej (klasyfikacji wielu etykiet) tych klas. Jest to zazwyczaj rozwiązywane w sposób nadzorowany. Przykładem klasyfikacji jest np. filtrowanie spamu, w której dane wejściowe to wiadomości e-mail (lub inne), a klasy to „spam” i „nie spam”.

K-najbliższych sąsiadów

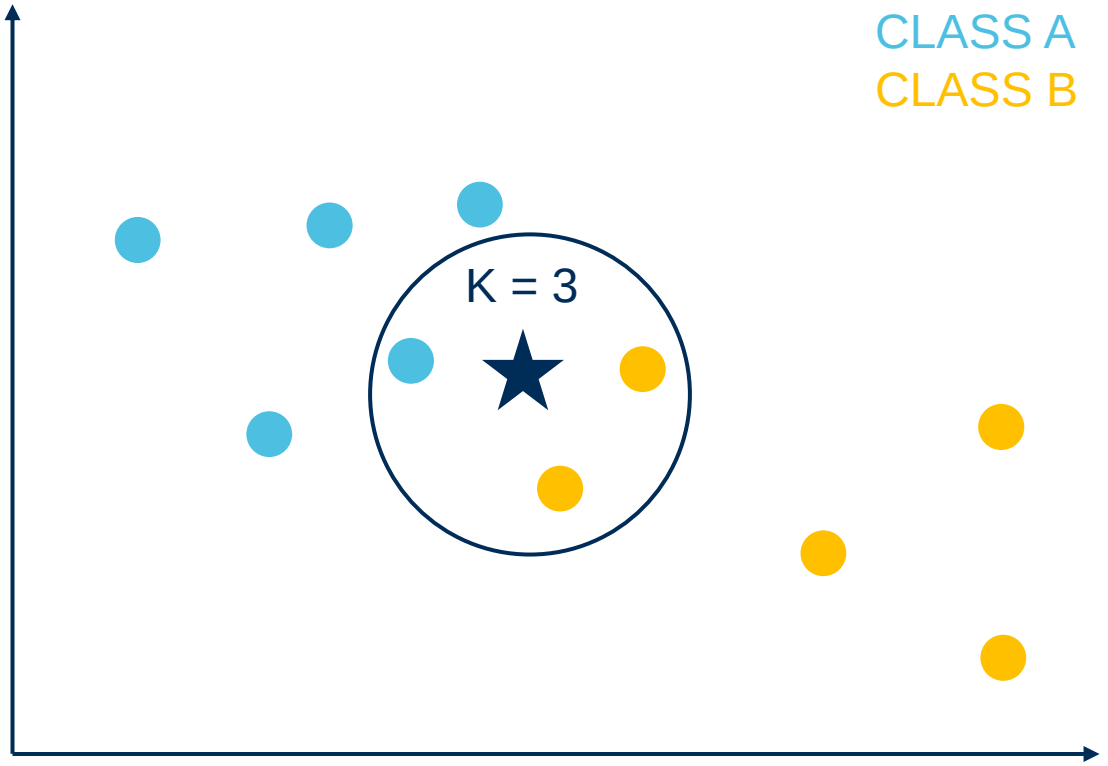
- Prosty klasyfikator (ściślej: algorytm regresji nieparametrycznej używany w statystyce do prognozowania wartości pewnej zmiennej losowej)
- Klasyfikacja nowych przypadków jest realizowana na bieżąco, tj. gdy pojawia się potrzeba klasyfikacji nowego przypadku
- Schemat algorytmu: Poszukaj obiektu/ów najbliższego w stosunku do obiektu klasyfikowanego, określ klasę nowego obiektu na podstawie klasy obiektu/ów najbliższego.

K-najbliższych sąsiadów

Zalety	Wady
Prosty w użyciu i implementacji	Posiada duże wymogi pamięciowe – musi przechować informacje o wszystkich przypadkach testowych w pamięci.
W łatwy sposób można wyjaśnić jak doszło do ustalenia (predykcji) klasy	Konieczność podania wartości k (liczby sąsiadów). Jest ona zależna od typu i specyfiki danych. Dla wartości k = 1, algorytm charakteryzuje się dużą podatnością na wstępowanie szumu informacyjnego.
Odporny na wartości izolowane – przez ocenę najbliższych sąsiadów	Czas dokonania klasyfikacji zwiększa się wraz z powiększaniem się zbioru danych, ponieważ zawsze trzeba wyliczyć odległość do wszystkich obiektów ze zbioru danych. Jest on zazwyczaj dużo dłuższy niż z użyciem modelu do klasyfikacji.

KNN - odległość

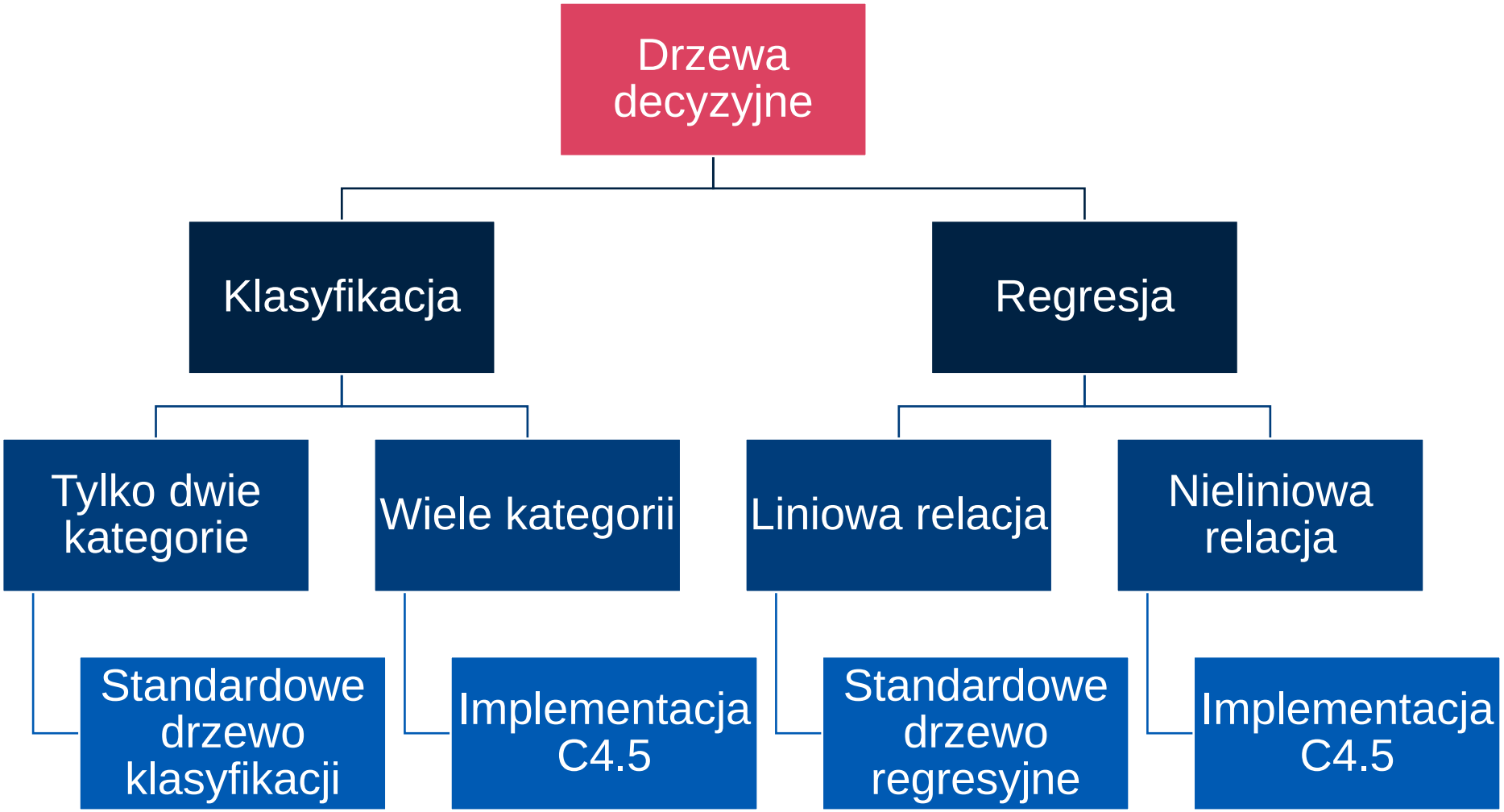
- Euklidesowa
- Minkowskiego
- Miejska (manhatańska)
- Czebyszewa



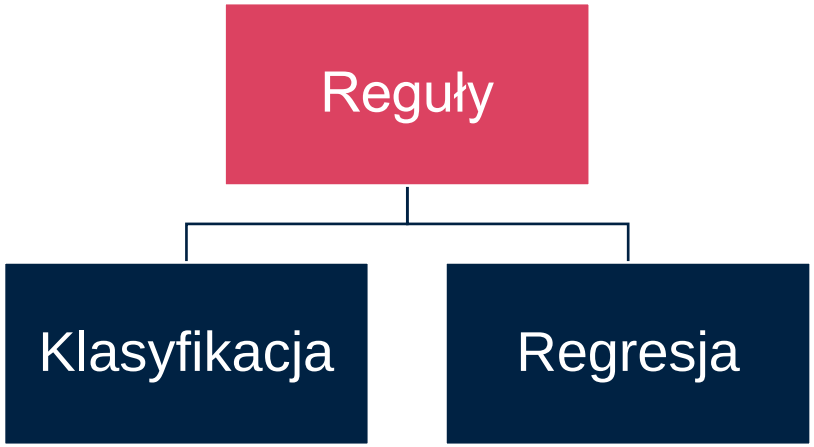
Drzewa decyzyjne

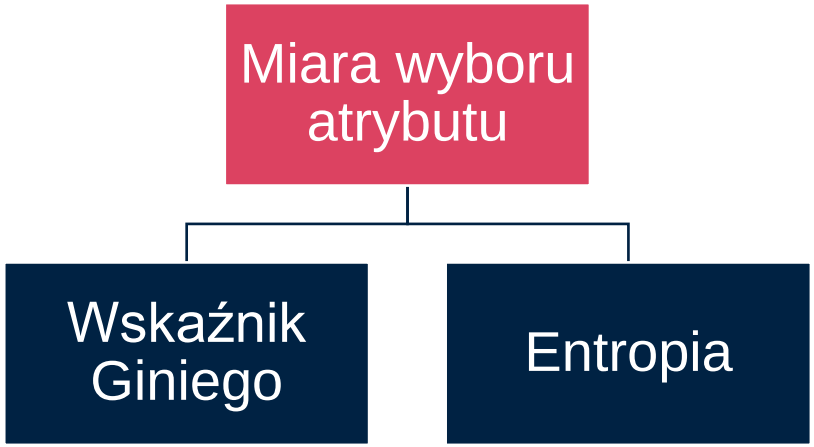
- Drzewa decyzyjne to graficzna metoda wspomagania procesu decyzyjnego
- Algorytm klasyfikacyjny zastosowany w celu znalezienia kroków decyzyjnych aby określić kategorię wyniku
- Atrybuty są wybierane przez tzw. algorytm wyboru cech, w kolejności mającej zmaksymalizować zysk informacyjny z danego węzła

Drzewa decyzyjne

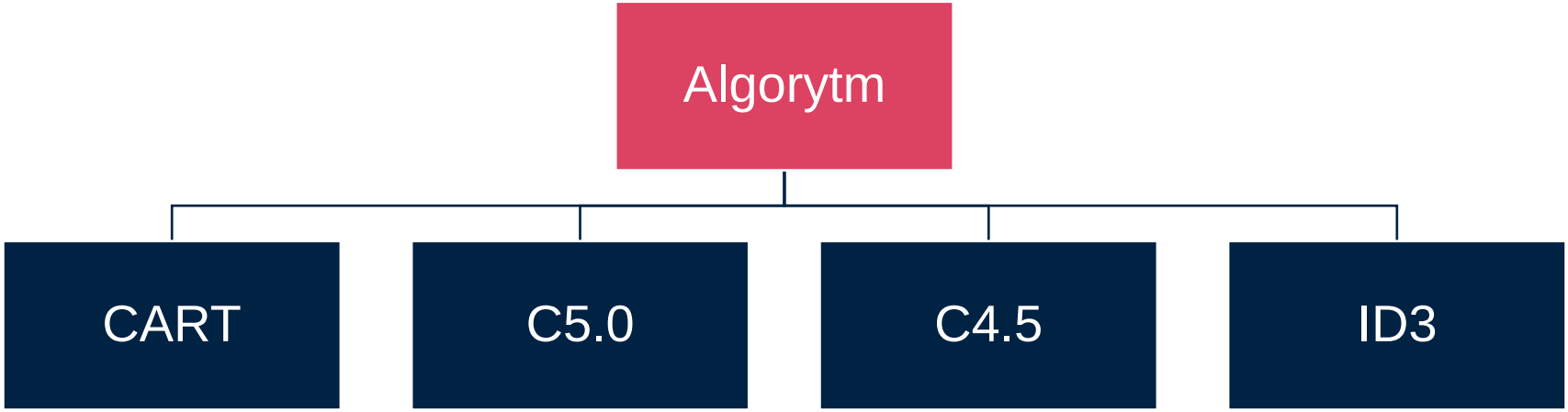


Drzewa decyzyjne



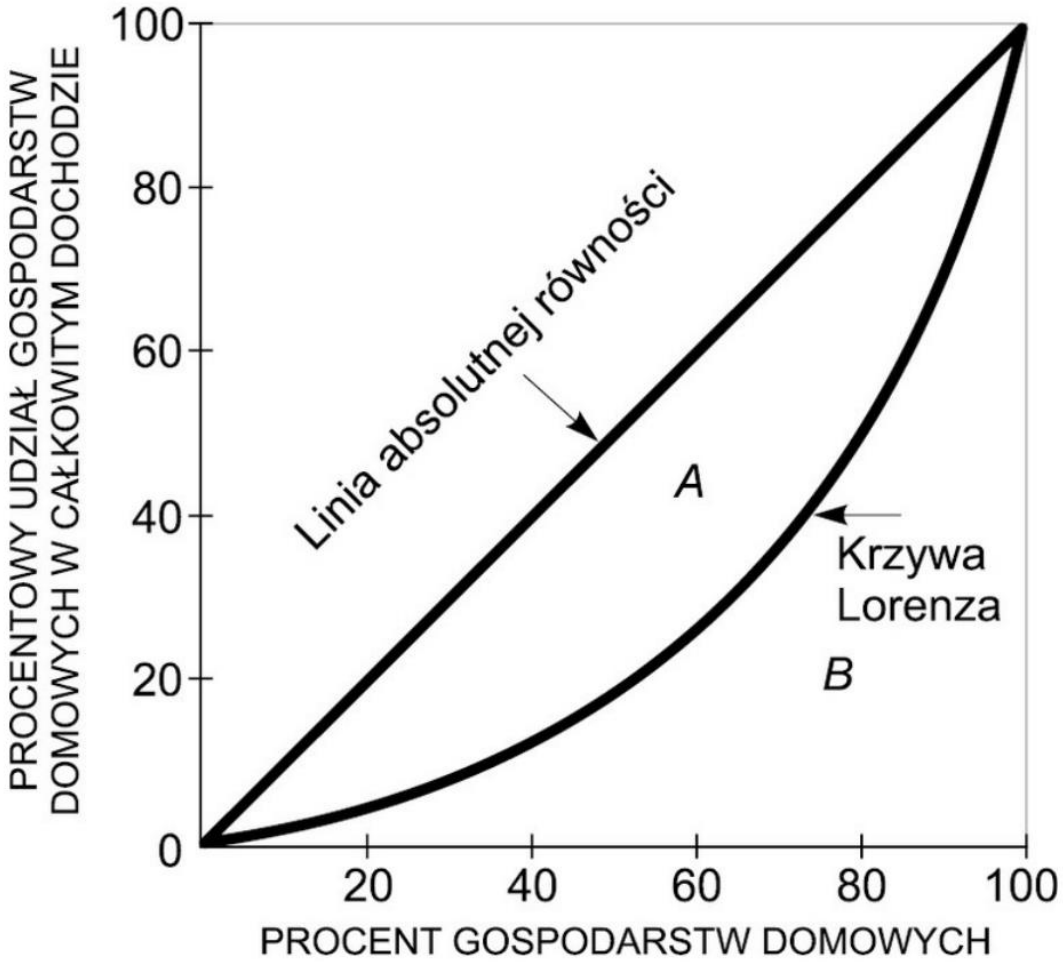


Drzewa decyzyjne



Gini indeks

$$G = \frac{\sum_{i=1}^{n}(2i - n - 1)x_i}{n\sum_{i=1}^{n} x_i}$$

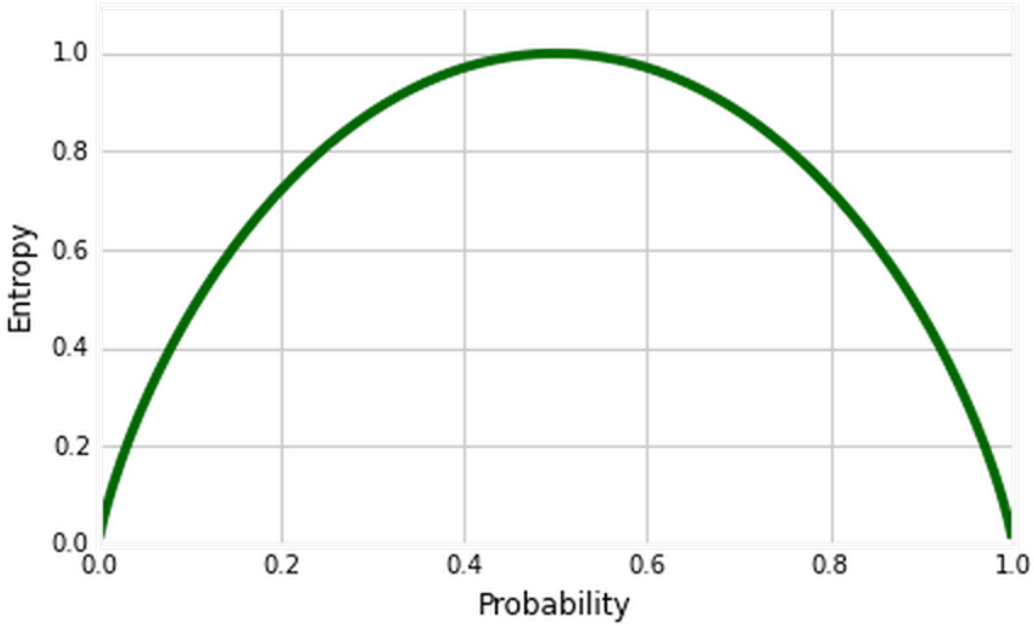


Stosowana w statystyce miara koncentracji (nierównomierności) rozkładu zmiennej losowej

Indeks (wskaźnik) Giniego

Entropia

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

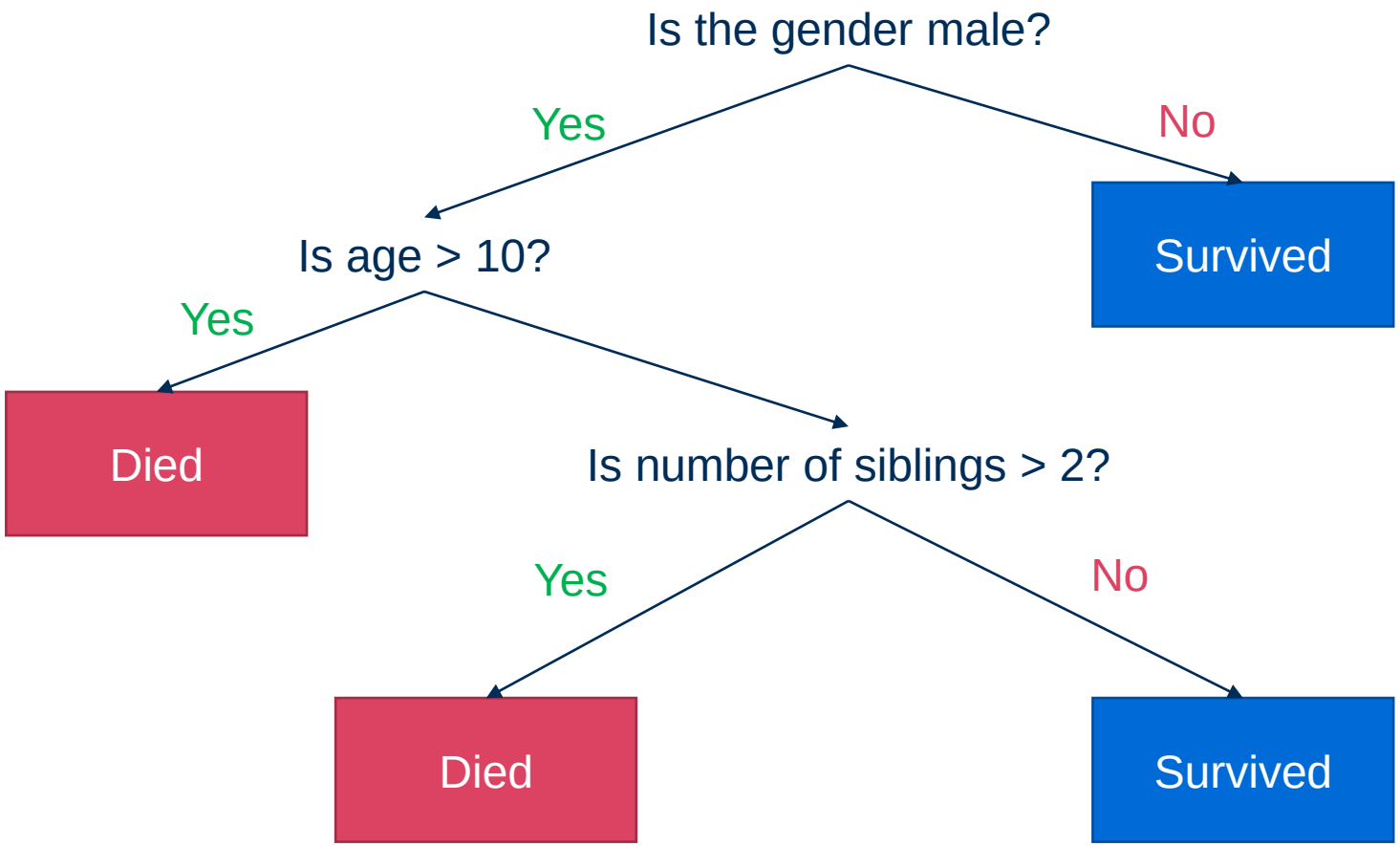


Miara niepewności (zanieczyszczenia) danych

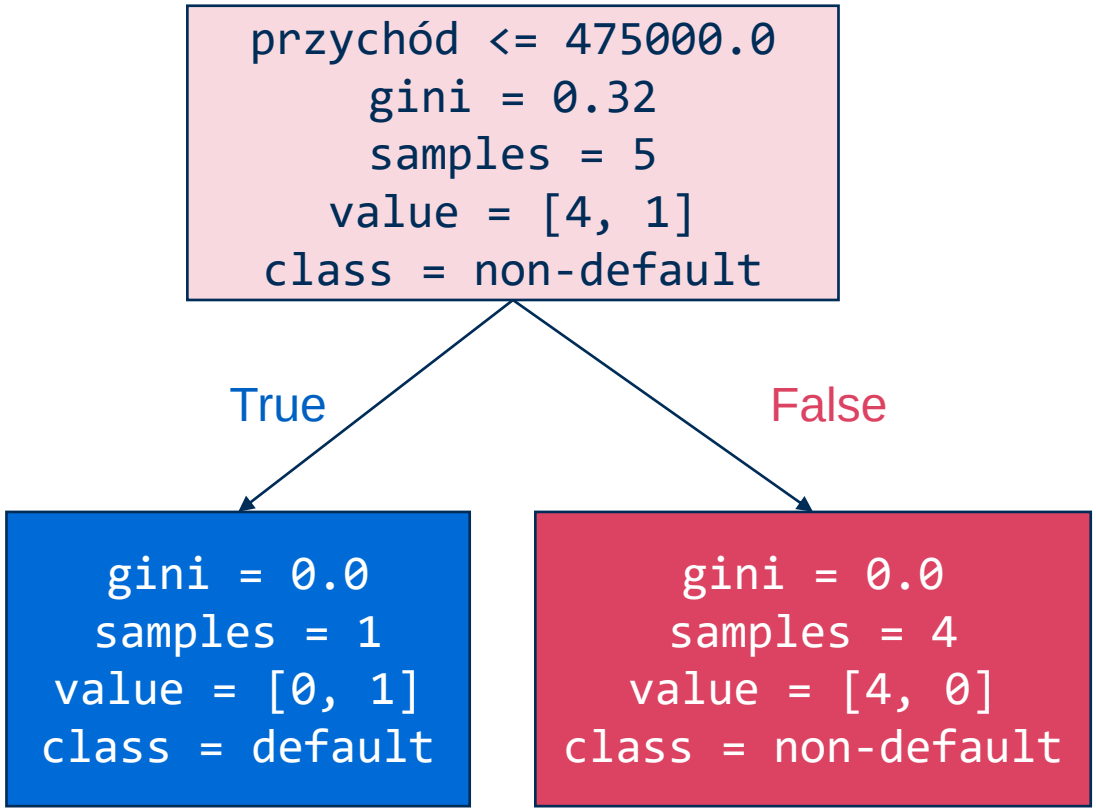
Drzewa decyzyjne

Zalety	Wady
uchwycenie zależności nieliniowych pomiędzy zmiennymi	**utrata części informacji** w procesie dyskretyzacji tracona jest część informacji
uwzględnienie braków danych w przypadku zmiennych kategorycznych nie trzeba usuwać niekompletnych obserwacji, gdyż brak można ująć jako odrębną kategorię (dotyczy to przypadku, w którym nowa zmienna zastępuje starą)	**czasem jej implementacja wymaga dodatkowego nakładu pracy** w przypadku niektórych sytuacji konieczne jest zapewnienie monotoniczności zmiennych, co powoduje, że w proces trzeba włożyć nieco więcej wysiłku
poprawa wyniku bez utraty interpretowalności modelu kategoryzacja jest jednym z zabiegów, po którego odpowiednim wykonaniu nie tracimy na interpretowalności modelu	

Drzewa decyzyjne

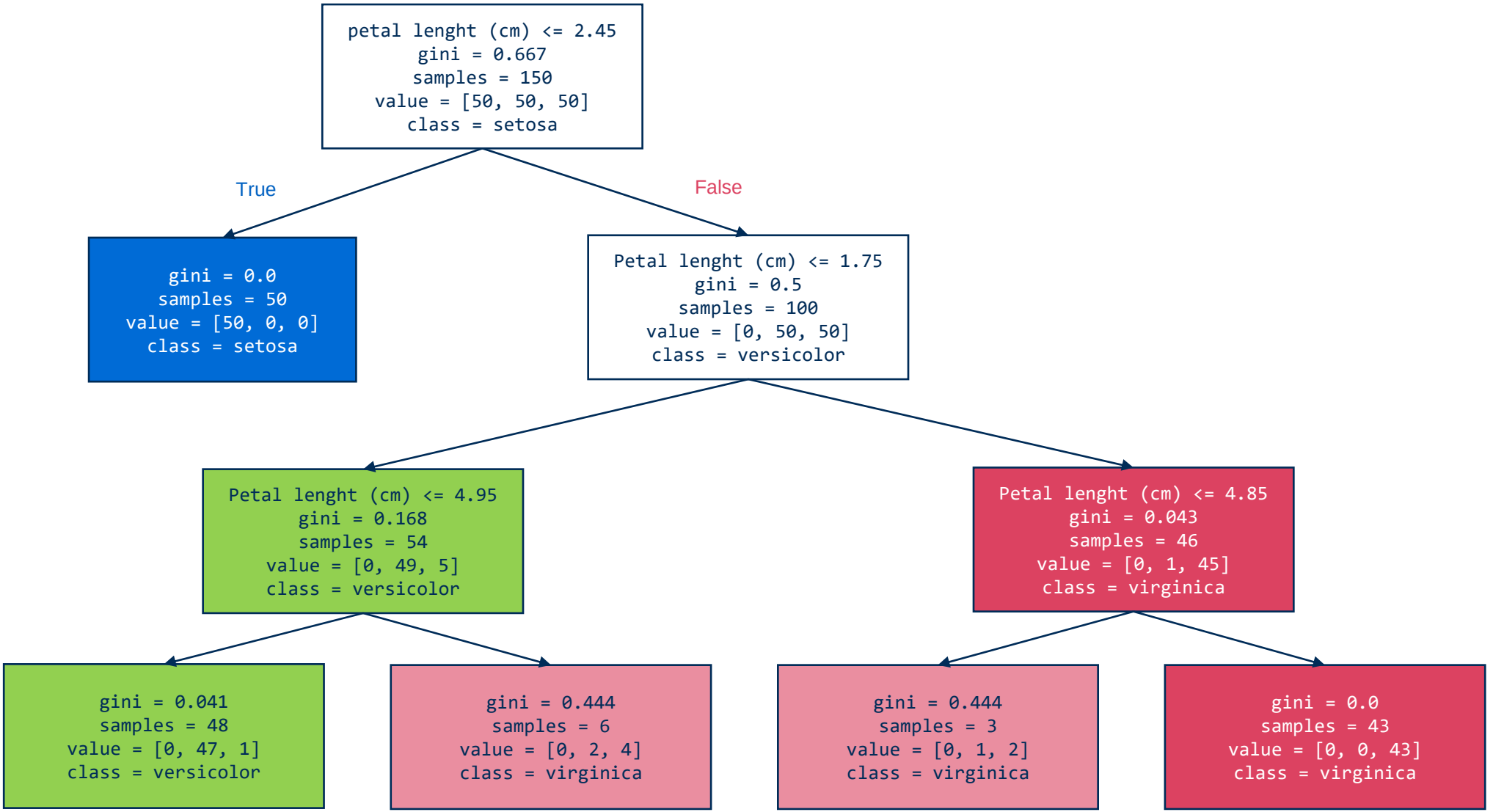


Drzewa decyzyjne

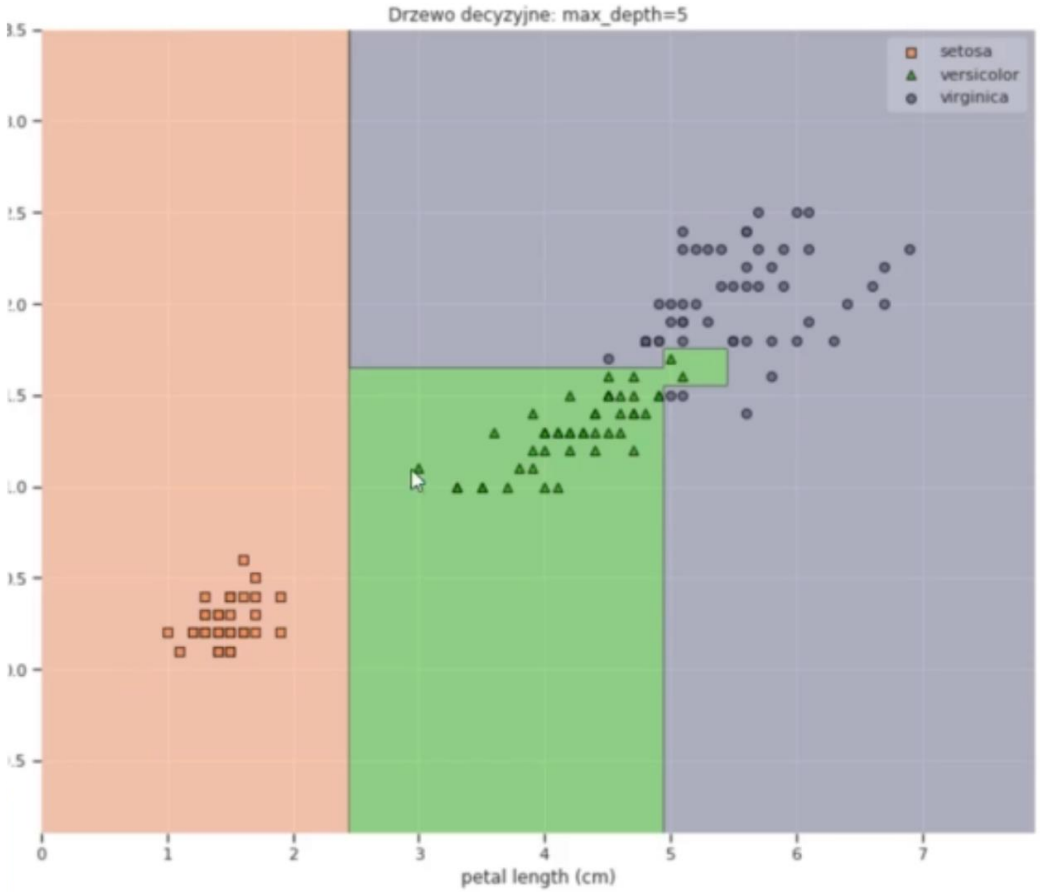


ID	Wiek	Przychód	Staż pracy	Default
1	25	60000	3	0
2	34	75000	12	0
3	27	35000	1	1
4	65	65000	40	0
5	37	85000	17	0

Drzewa decyzyjne



Drzewa decyzyjne

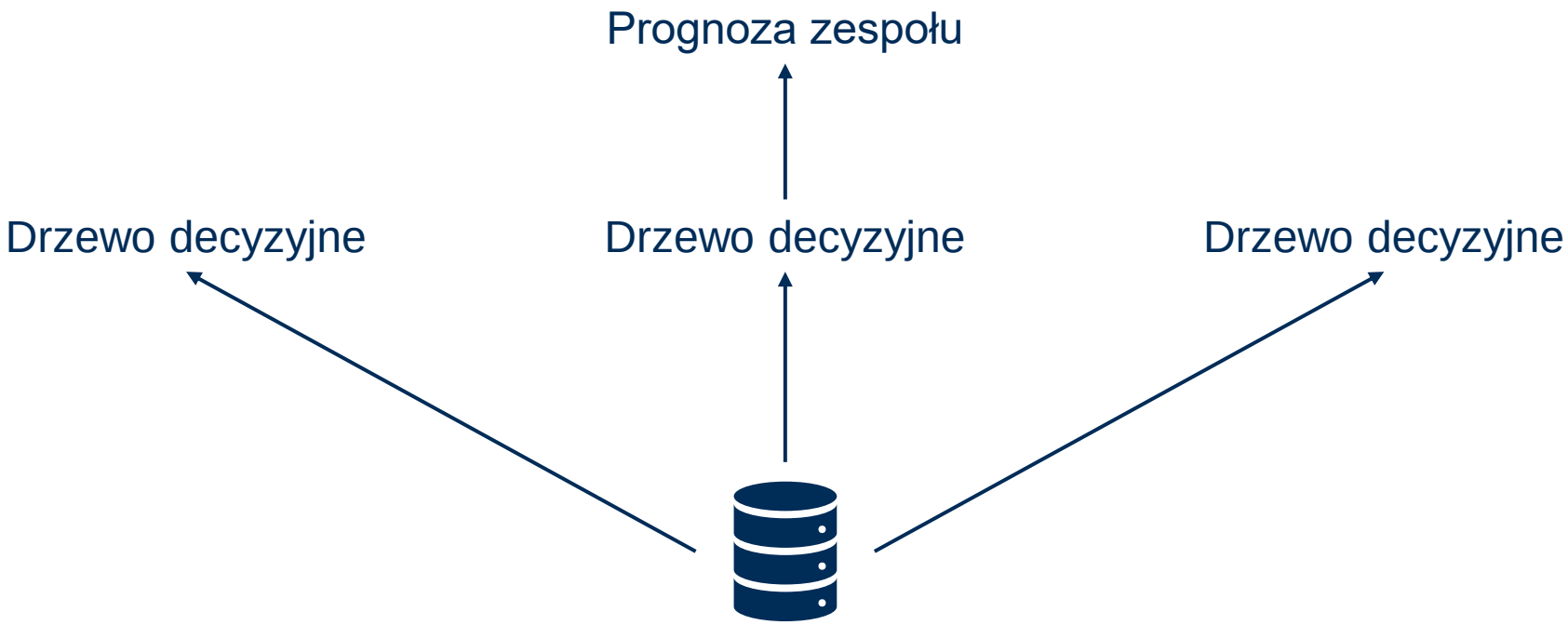


Lasy losowe

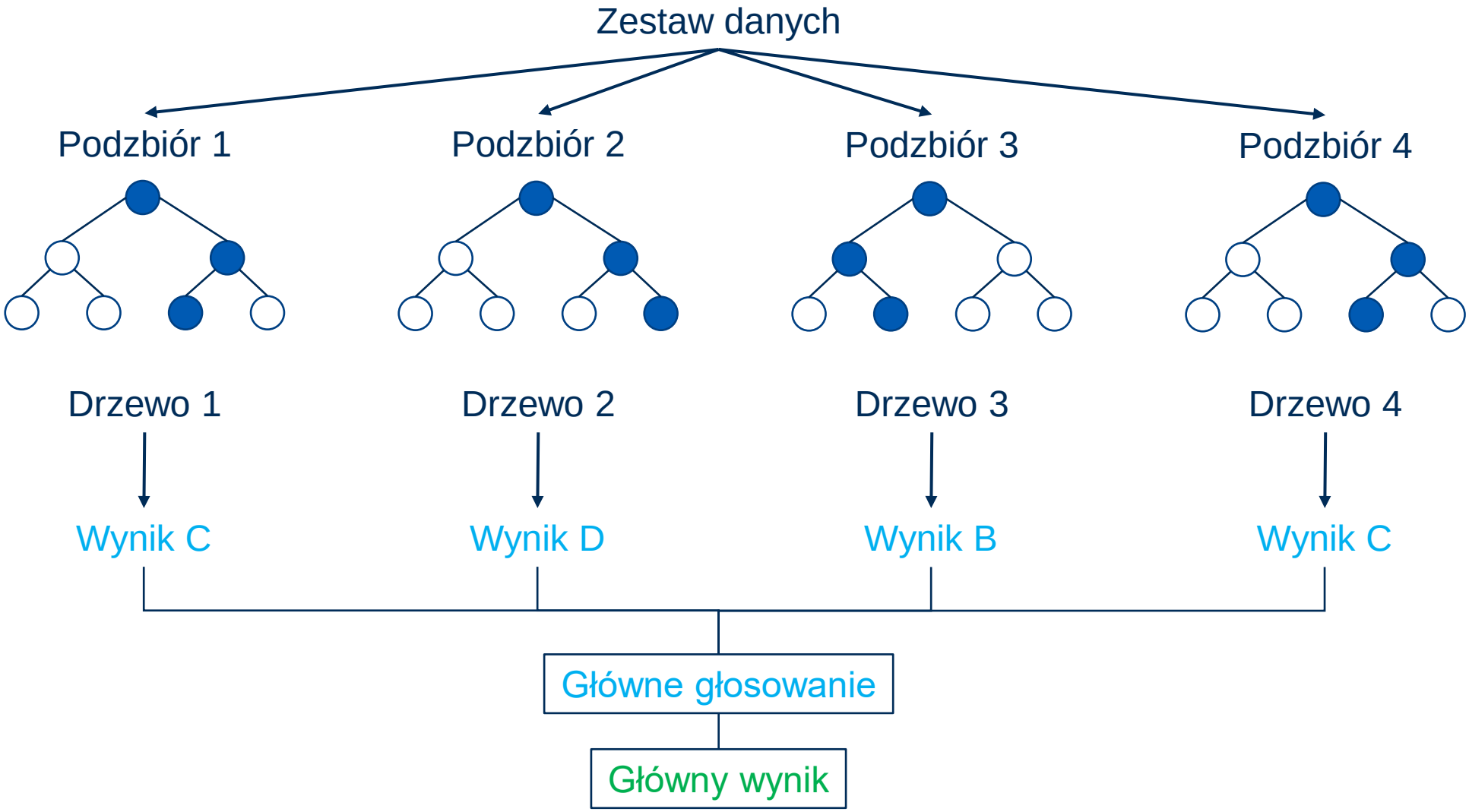
- Las losowy jest zbiorem drzew klasyfikacyjnych o podziałach binarnych
- Algorytm lasów losowych tworzy na zbiorze treningowym określoną przez użytkownika liczbę drzew klasyfikacyjnych. Aby zwiększyć przewagę lasu nad pojedynczym drzewem, pożądane jest zróżnicowanie (wariancja) drzew wchodzących w jego skład.

  Wariancja ta jest osiągana na następujące sposoby:
  - Losowy wybór zbioru treningowego dla każdego drzewa
  - Losowy lub deterministyczny wybór zmiennych, na których dokonane zostaną cięcia (splits), zarówno dla całego drzewa jak i pojedynczego cięcia
  - Losowy wybór metody cięcia dla każdego drzewa.
  - Losowe zróżnicowanie minimalnego rozmiaru węzła (w zadanych granicach) dla każdego drzewa.

Lasy losowe


Prognoza zespołu
Drzewo decyzyjne
Drzewo decyzyjne
Drzewo decyzyjne

Lasy losowe


Zestaw danych
Podzbiór 1
Podzbiór 2
Podzbiór 3
Podzbiór 4
Drzewo 1
Drzewo 2
Drzewo 3
Drzewo 4
Wynik C
Wynik D
Wynik B
Wynik C
Główne głosowanie
Główny wynik

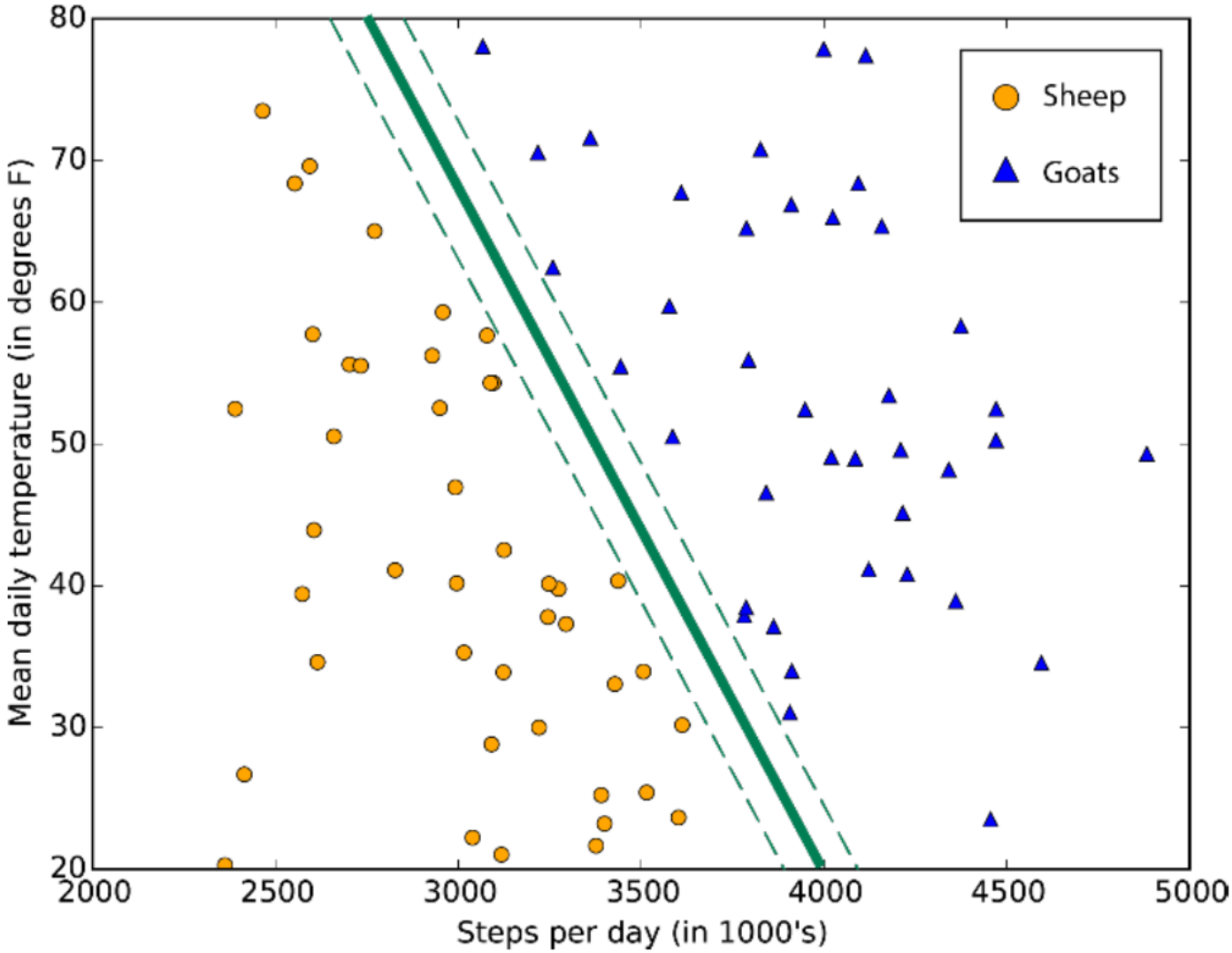
Funkcja straty

- Funkcja straty to funkcja przyporządkowująca nieujemną wielkość kary poprzez porównanie prawdy (założymy chwilowo, że ją znamy) do podjętej decyzji (wyliczonego estymatora)

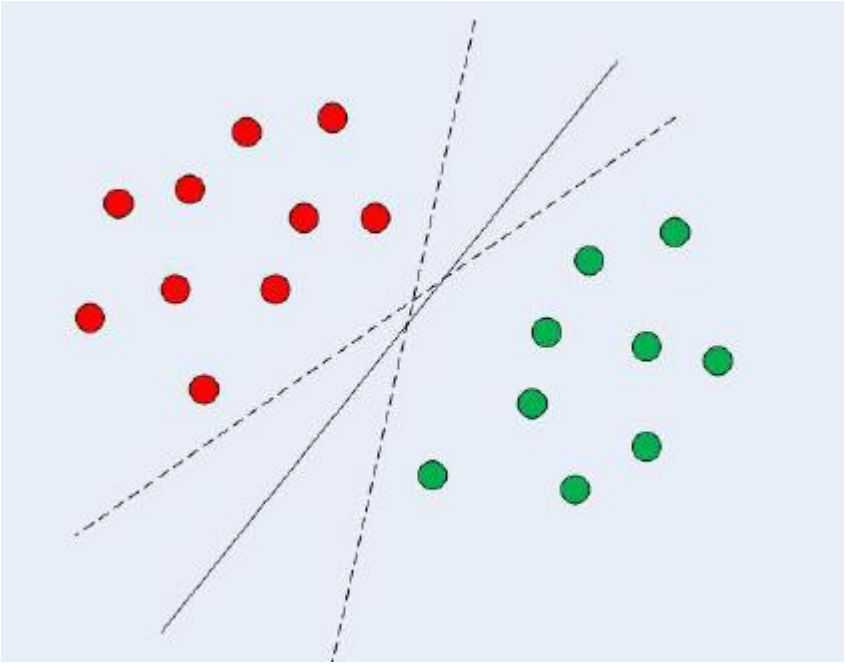
SVM - Support Vector Machine

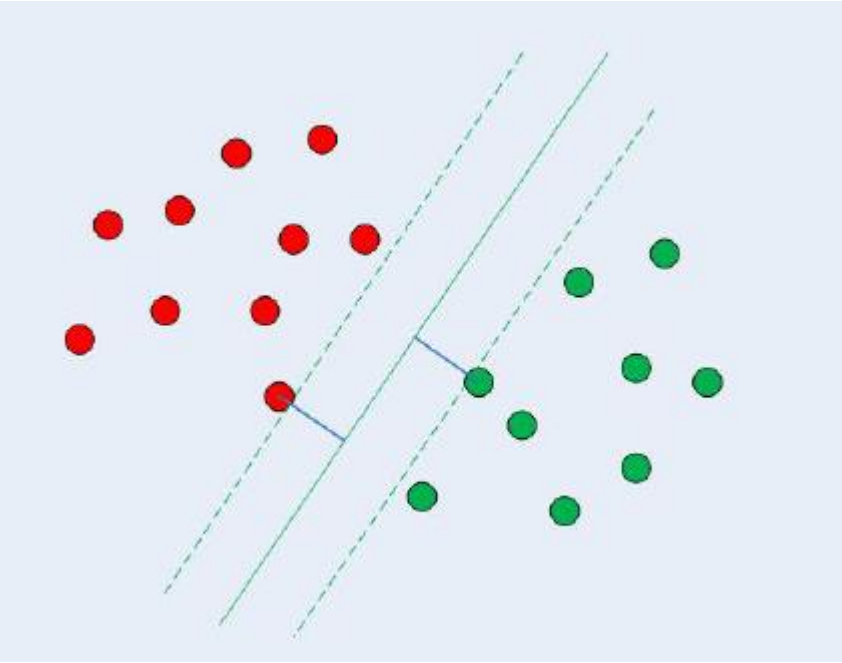
- Algorytm SVM to technika klasyfikacji i regresji maksymalizująca dokładność predykcyjną modelu bez przeuczenia danych uczących. Algorytm SVM szczególnie dobrze nadaje się do analizowania danych o bardzo dużej liczbie (np. tysiącach) zmiennych predykcyjnych.
- Algorytm SVM znajduje zastosowanie w wielu dziedzinach, takich jak zarządzanie relacjami z klientami (CRM), rozpoznawanie twarzy i innych obrazów, bioinformatyka, rozpoznawanie głosu i mowy.
- Działanie algorytmu SVM polega na mapowaniu danych na wielowymiarową przestrzeń właściwości w sposób umożliwiający kategoryzację punktów danych, nawet jeśli danych tych nie można w inny sposób liniowo oddzielić
- Najpierw odszukiwany jest separator między kategoriami. Następnie dane są przekształcane w sposób umożliwiający wyrysowanie separatora jako hiperpłaszczyzny
- Po wykonaniu tych czynności charakterystyki nowych danych mogą służyć do przewidywania grupy, do której powinien należeć nowy rekord

SVM - Support Vector Machine



SVM - Support Vector Machine

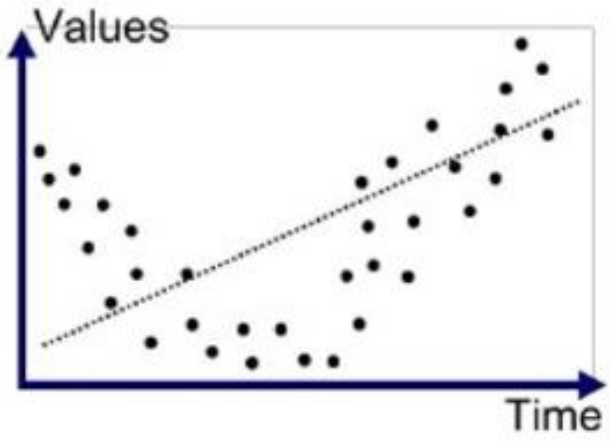




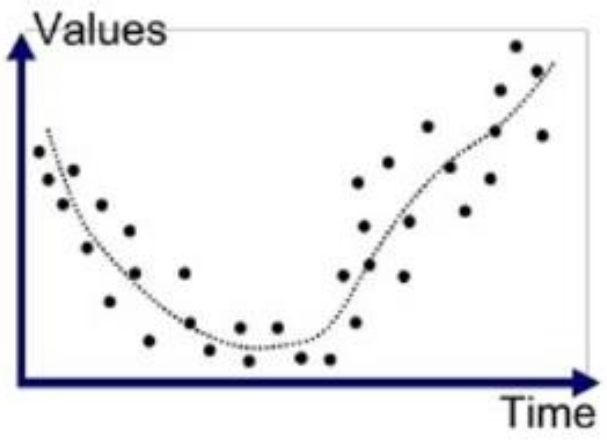
SVM - Support Vector Machine

- **Kernel**
  funkcja przekształcająca obszar nierozdzielny do rozdzielnego
  - Liniowa
  - Wielomianowa
  - Radialna funkcja bazowa (RBF)
  - Sigmoidalna (zwana też krzywą logistyczną)
- **Regularyzacja**
  C parametr używany do utrzymywania regularyzacji
- **Gamma**
  niższa wartość Gammy będzie luźno pasować do zbioru danych treningowych, podczas gdy wyższa wartość gamma będzie dokładnie pasować do zbioru danych treningowych, co powoduje nadmierne dopasowanie

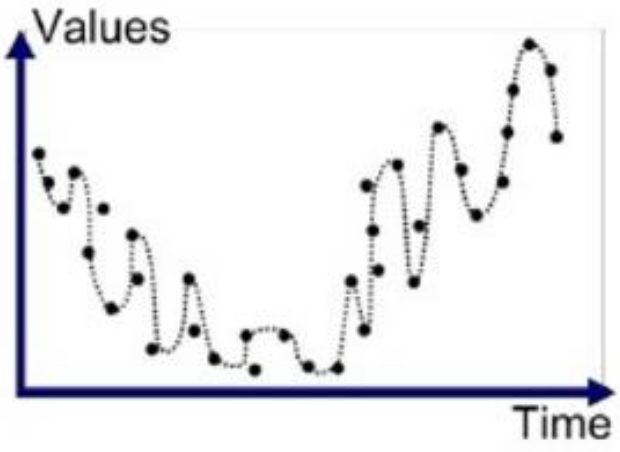
Przeuczenie/niedouczenie



Niedouczony



Dobrze dopasowany (solidny)

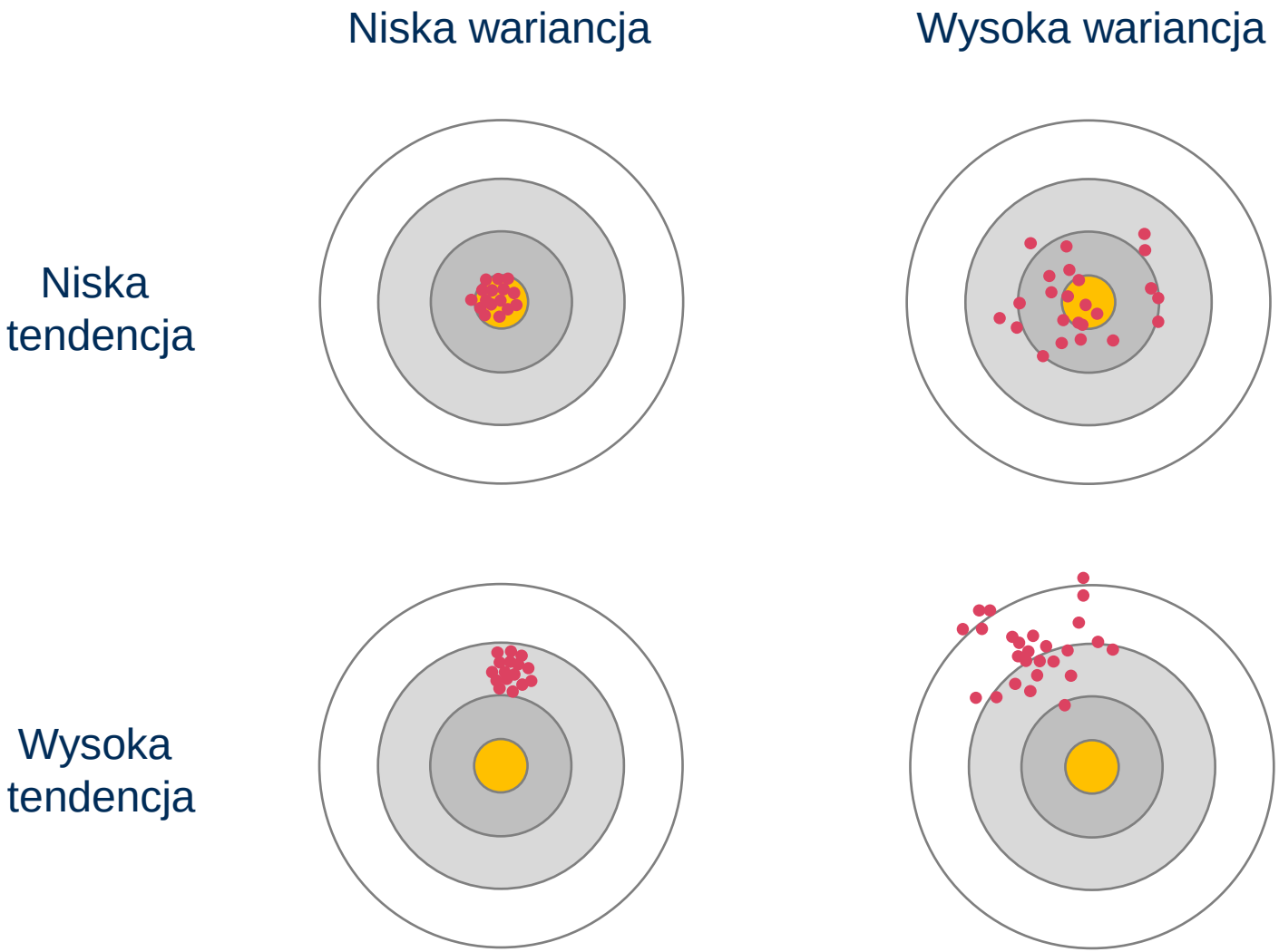


Przeuczony

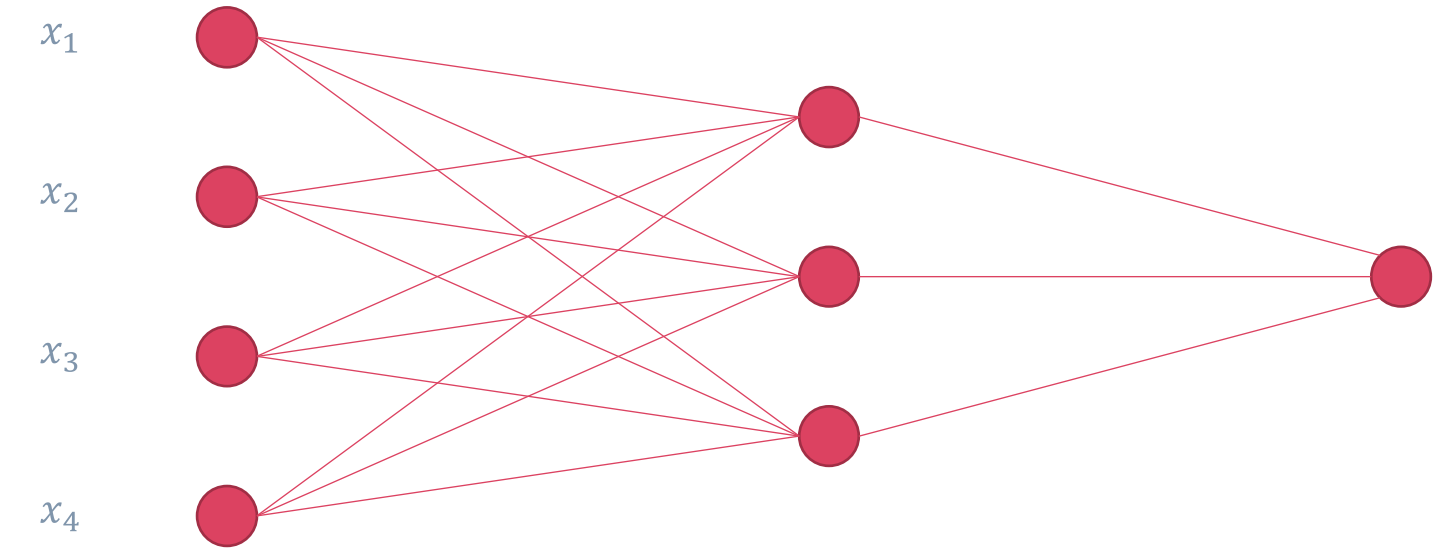
Przeuczenie/niedouczenie

- **Niedouczenie**
  - Upewnij się, że istnieje wystarczająca ilość danych treningowych, aby funkcja błędu / kosztu (np. MSE lub SSE) była wystarczająco zminimalizowana
- **Nadmierne dopasowanie**
  - Ogranicz liczbę funkcji lub regulowane parametry w modelu. Wraz ze wzrostem liczby funkcji wzrasta również złożoność modelu, co zwiększa szansę na nadmierne dopasowanie
  - Skróć szkolenie, aby model nie „nadmiernie nauczył się” danych treningowych
  - Dodaj formę regularyzacji do funkcji błędu / kosztu, aby zachęcić do płynniejszego odwzorowania (często stosuje się regresję Ridge lub Lasso)

Przeuczenie/niedouczenie



Sieci Neuronowe - warstwy



Warstwa wejściowa $\in \mathbb{R}^4$ *Warstwa ukryta* $\in \mathbb{R}^3$ *Warstwa wyjściowa* $\in \mathbb{R}^1$

Sieci Neuronowe - warstwy

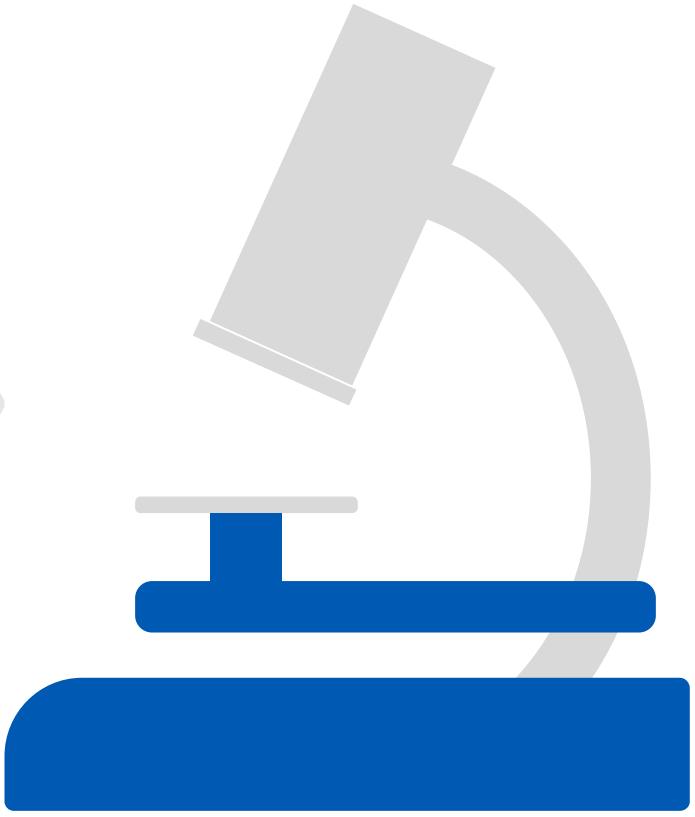
- **Dense** – sieć gęsta
- **Activation** – stosuje funkcję aktywacji
- **Dropout** – porzuca wskazaną liczbę neuronów
- **Flatten** – wypłaszcza dane wejściowe
- **Input** – warstwa wejściowa

Sieci Neuronowe – rodzaje sieci

Sieci jednowarstwowe

Sieci wielowarstwowe

Sieci rekurencyjne



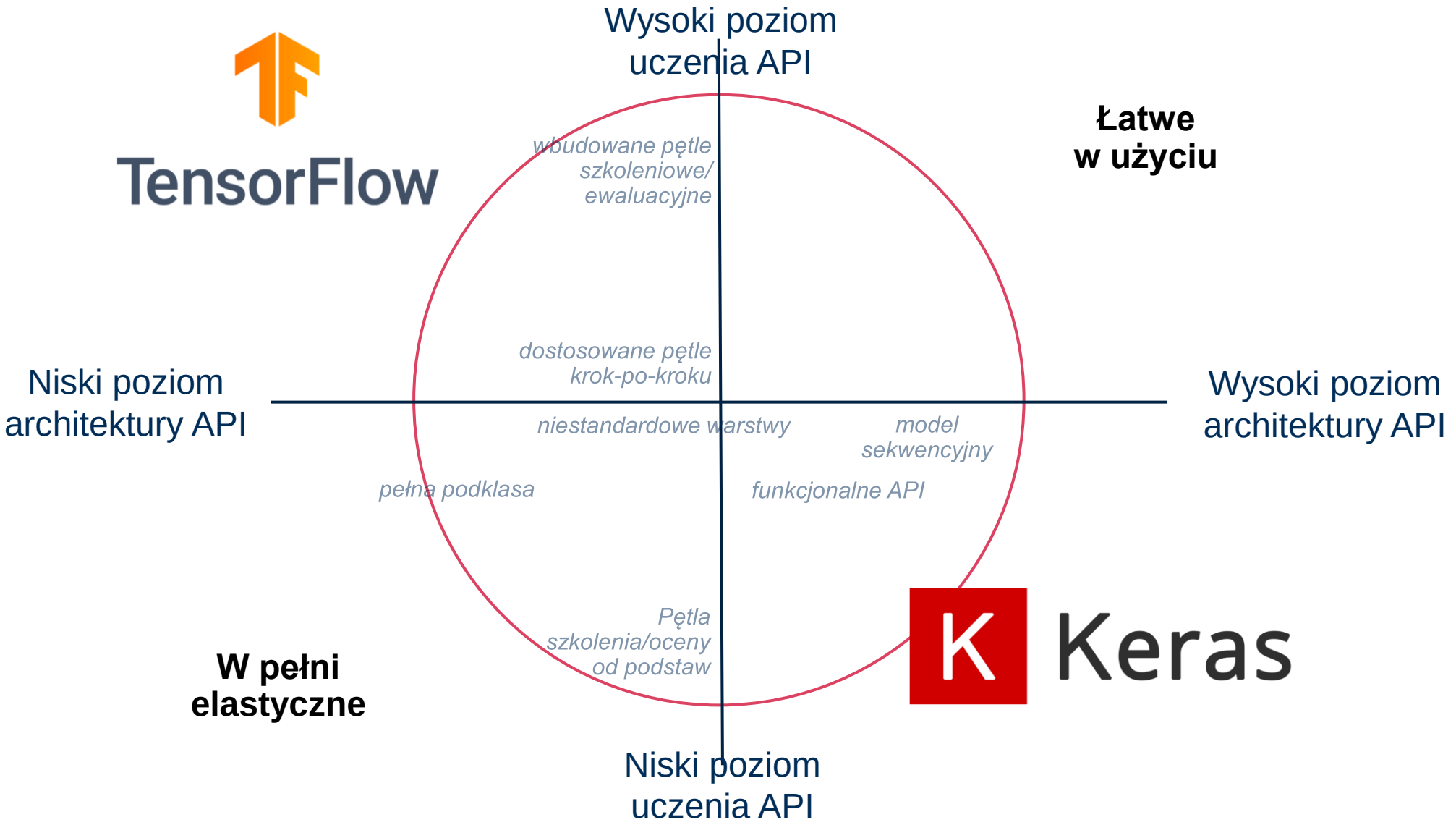
Sieci Neuronowe – rodzaje sieci

- Sieci **jednowarstwowe** - neurony w tej sieci ułożone są w jednej warstwie, zasilanej jedynie z węzłów wejściowych. Węzły wejściowe nie tworzą warstwy neuronowej, ponieważ nie zachodzi w nich proces obliczeniowy. Sieci tego rodzaju mogą stanowić zarówno perceptrony jednowarstwowe, jak i sieci Kohonena.
- Sieci **wielowarstwowe** - ich cechą charakterystyczną jest występowanie co najmniej jednej warstwy ukrytej neuronów, pośredniczącej w przekazywaniu między węzłami wejściowymi a warstwą wyjściową. Neurony warstw ukrytych stanowią bardzo istotny element sieci. Często sieci wielowarstwowe jednokierunkowe nazywamy perceptronami wielowarstwowymi.
- Sieci **rekurencyjne** - najczęściej sieci rekurencyjne składają się z jednej warstwy neuronów np. Sieć Hopfielda. Występuje w nich sprzężenie zwrotne między warstwami wyjściową i wejściową. Sygnały wyjściowe neuronów tworzą jednocześnie wektor wejściowy sieci dla następnego cyklu.

Sieci Neuronowe – architektura

Sieci Neuronowe – biblioteki

Sieci Neuronowe – biblioteki



Sieci Neuronowe – funkcje aktywacji

- Funkcja aktywacji – pojęcie używane w sztucznej inteligencji do określenia funkcji, według której obliczana jest wartość wyjścia neuronów sieci neuronowej:
  - ReLU
  - Sigmoid
  - Tanges hiperboliczny (tanh)
  - Softmax

Sieci Neuronowe – funkcje straty

$$Binary\ CrossEntropy = -\frac{1}{N}\sum_{i=1}^{N}(y_i * log(p(y_i)) + (1 - y_i) * log(1 - p(y_i))$$

$$Categorical\ CrossEntropy = -\sum_{i} y_i * log(p(y_i))$$

$$MAE = \frac{1}{n}\sum_{i=1}^{N}|y_{true} - y_{pred}|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_{true} - y_{pred})^2$$

Sieci Neuronowe – optymalizator

SGD

RMSprop

Adagrad

Adadelta

Adam

Adamax

Nadam

Dziękuję!