Charakterystyka wybranych modeli uczenia maszynowego

Adam Wrzałek, 294485

Mikołaj Mazur, 294482

Bartosz Deptuła, 254010

Różnorodność modeli uczenia maszynowego

Rodzaje modeli uczenia maszynowego	Rodzaje w zależności od zadania	Rodzaje w zależności od struktury danych	Podział na proste i złożone	Podział ze względu na różne podejścia do modelowania	Podział ze względu na przykłady praktycznych zastosowań
Nadzorowane (regresja liniowa, drzewa decyzyjne, sieci neuronowe)	Klasyfikujące (kNN, random forest)	Modele danych tabelarycznych (XGBoost, LightGBM)	Modele proste (regresja logistyczna, regresja liniowa)	Modele probabilistyczne (HMM, modele Bayesowskie)	Modele do przewidywania szeregów czasowych (ARIMA, LSTM)
Nienadzorowane (analiza skupień, algorytm K-średnich)	Regresyjne (regresja wielomianowa)	Modele przetwarzania języka naturalnego (LSTM, GPT)	Modele złożone (głębokie sieci neuronowe, AdaBoost)	Modele oparte na uczeniu głębokim (RNN, GAN)	Modele do detekcji anomalii (Isolation Forest)
Ze wzmocnieniem (Q-learning)	Klastrowe (DBSCAN, hierarchiczna analiza skupień)	Modele przetwarzania obrazu (CNN, ResNet)		Modele oparte na drzewach (drzewa decyzyjne)	
	Generujące (GAN, modele sekwencyjne)				

Regresja Liniowa

- *Jest to jedna z najbardziej podstawowych i przydatnych metod analizy danych.*
- *Wykorzystywana jest powszechnie w uczeniu maszynowym w celu wyznaczania relacji pomiędzy zmiennymi.*

Zasada działania

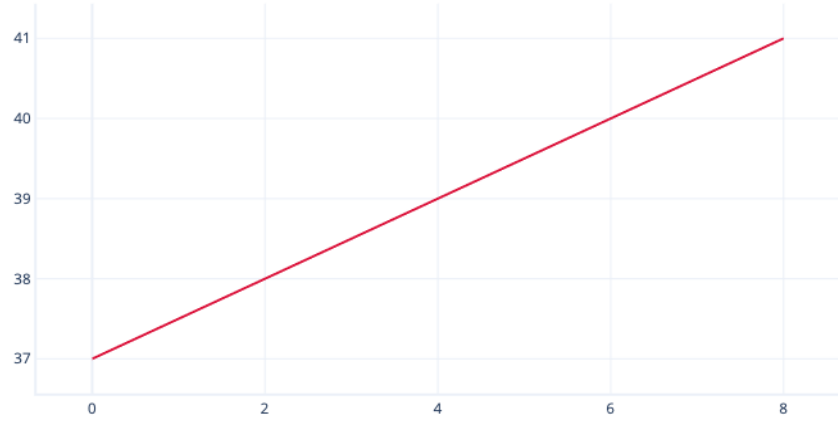
Regresja liniowa modeluje relację pomiędzy funkcją a etykietą ciągłą.

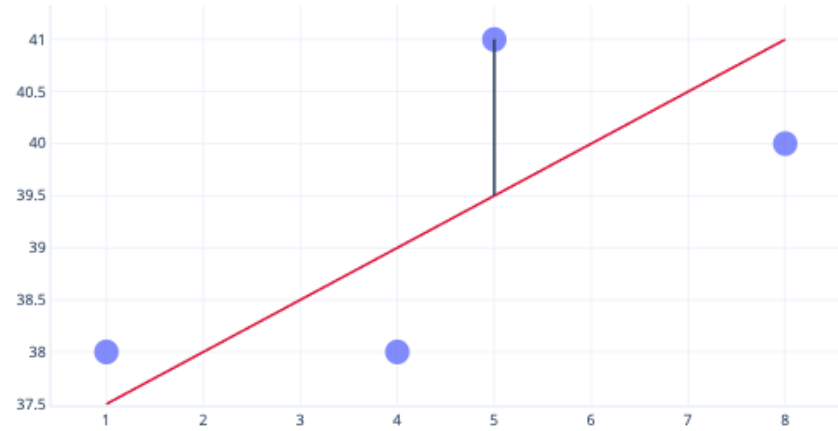
Postać funkcji: y=m*x + c

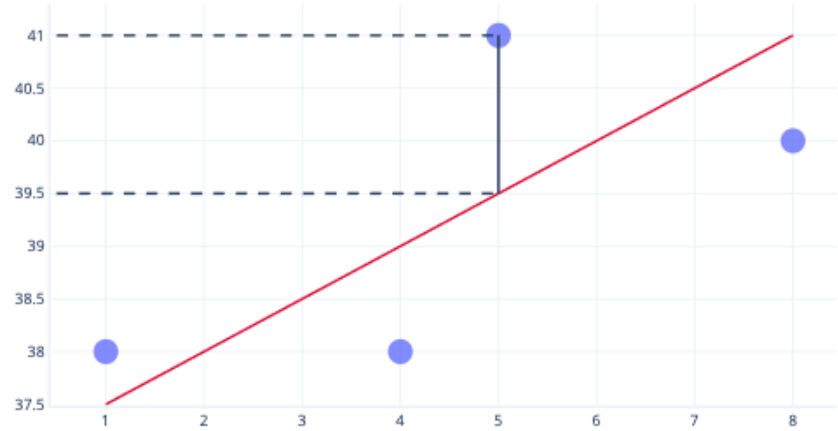
Posiada ona dwa parametry: przecięcie *(c)* i nachylenie *(m)*.

Podczas tworzenia dopasowuje się model regresji do danych tak, aby wygenerować jak najmniej błędów, czyli różnicy pomiędzy rzeczywistą a przewidywaną wartością.

Najczęściej dopasowanie do modelu polega na minimalizowaniu sumy reszt kwadratów.







Zalety regresji liniowej

Wady regresji liniowej

Zakłada liniową zależność między zmiennymi

Bardzo wrażliwa na wartości odstające

Nie będzie pasować do znacznej części danych

Regresja Logistyczna

- *Jest to metoda pozwalająca na modelowanie szansy i prawdopodobieństwa wystąpienia zdarzenia.*
- *Wykorzystywana jest powszechnie w uczeniu maszynowym w celu podziału na dwie klasy.*

Zasada działania

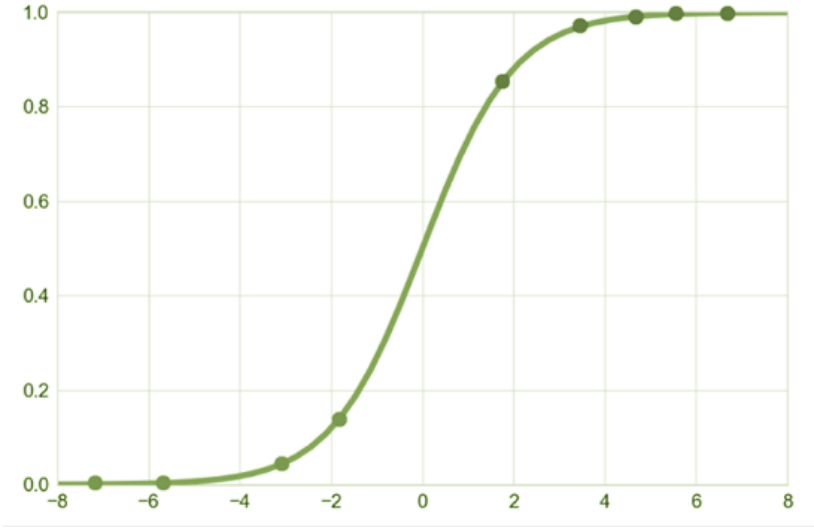
Prawdopodobieństwo jest to ilość obserwowanych zajść zdarzenia w całkowitej próbie.

Szansa jest to iloraz prawdopodobieństwa wystąpienia zdarzenia i jego niewystąpienia.

Wzór występuje w postaci logarytmu naturalnego szansy, gdzie (a) to wyraz wolny, (b) to współczynnik regresji logistycznej a (x) to zmienna objaśniająca

Model analizuje związek pomiędzy jedną lub wieloma zmiennymi niezależnymi i klasyfikuje dane do klas dyskretnych

Wynikiem regresji logistycznej będzie określenie czy zmiana wartości zmiennej objaśniającej przewiduje mniejsze, czy większe prawdopodobieństwo zajścia zdarzenia



$$\ln\left(\frac{P_i}{1-P_i}\right) = a + bx$$

Zalety regresji logistycznej

Łatwa do wdrożenia metoda uczenia maszynowego

Odpowiednia dla liniowo rozdzielnych zestawów danych

Miara istotności zmiennej predykcyjnej oraz określenia jej relacji i powiązania

Wysoka dokładność w przypadku wielu prostych zbiorów danych

Można ją łatwo rozszerzyć na wiele klas

Wady regresji logistycznej

Drzewo decyzyjne

Cel

Drzewo decyzyjne, dzieli dostępne dane na coraz mniejsze podzbiory na podstawie wartości wybranych cech. Każdy podział zmniejsza niepewność co do tego, do jakiej klasy należy dana obserwacja.

Budowa

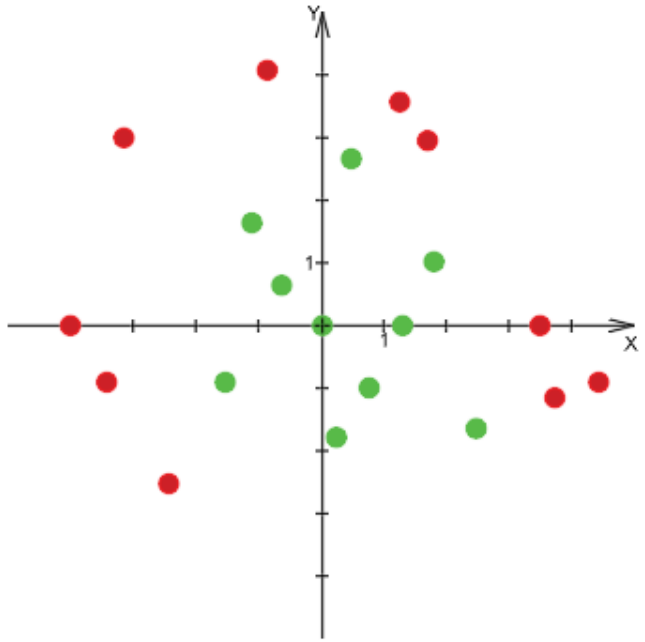
węzeł główny – to pierwszy węzeł drzewa. Ten węzeł zawiera wszystkie dane wejściowe i od niego wychodzi pierwsze pytanie, które dzieli nasz zbór na podzbiory.

węzeł decyzyjny – Węzły te reprezentują pośrednie decyzje w drzewie.

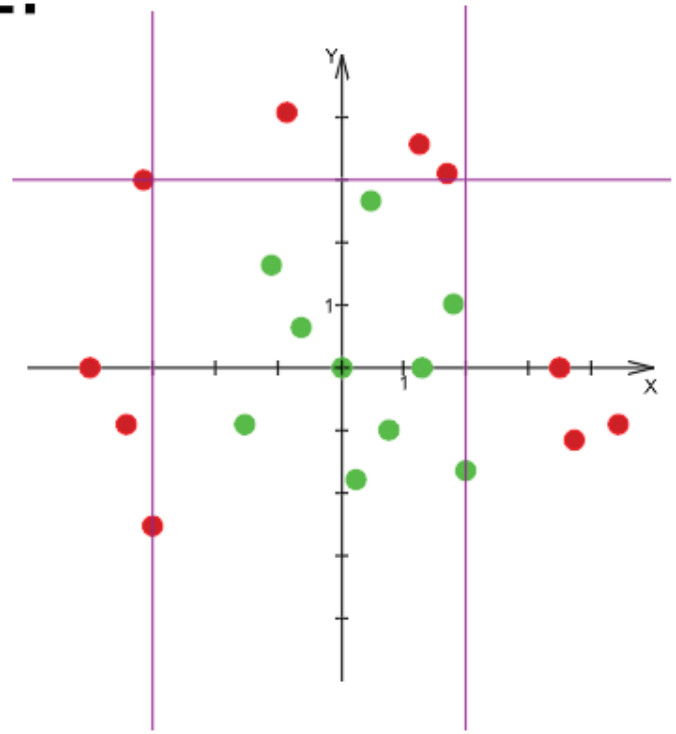
węzeł liściowy (węzeł końcowy) – jest to ostatni węzeł, który nie ulega dalszemu podziałowi.

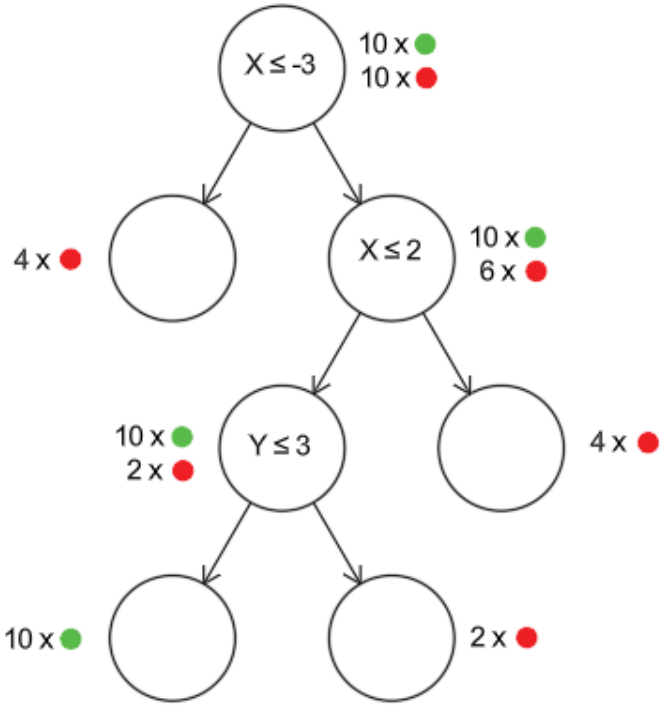
Zasada działania

1.

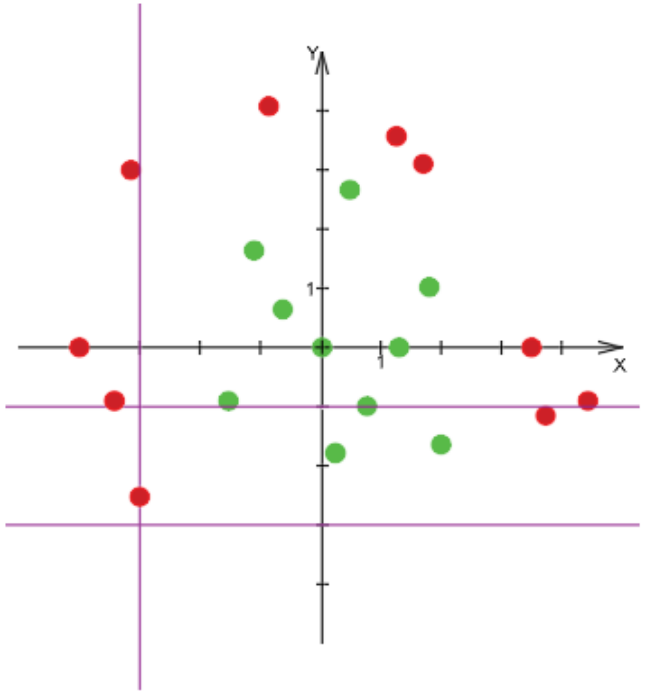


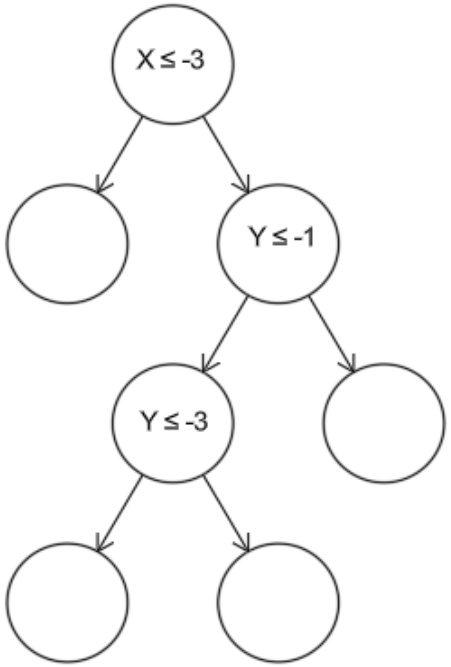
2.

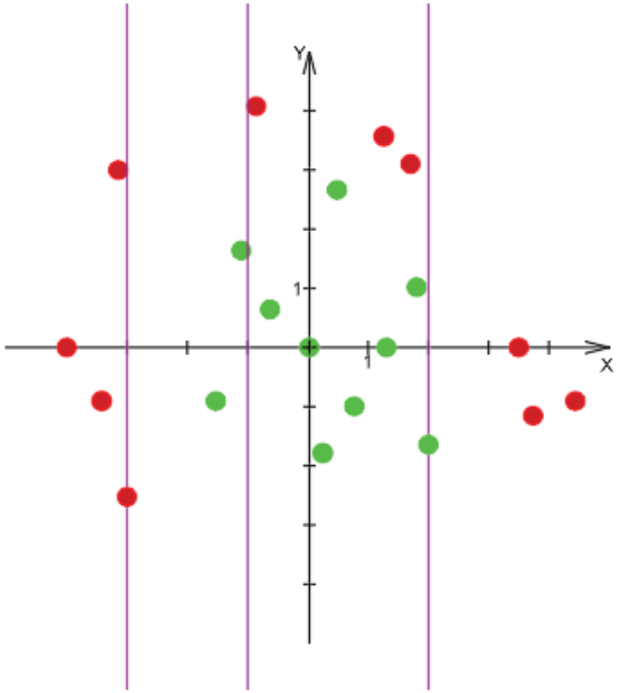


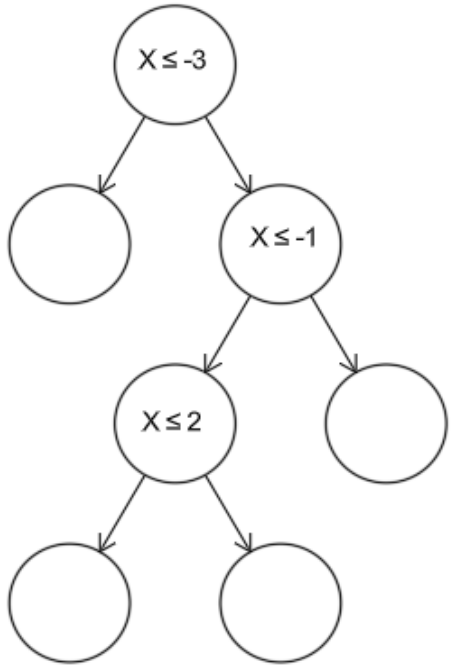


Inne możliwości podziału:

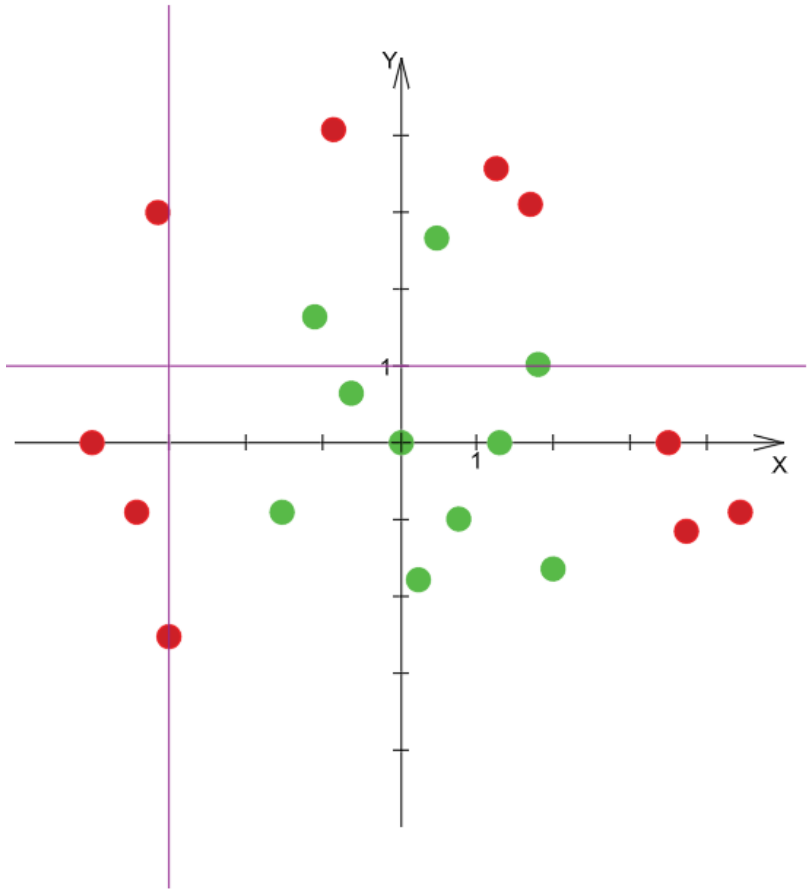






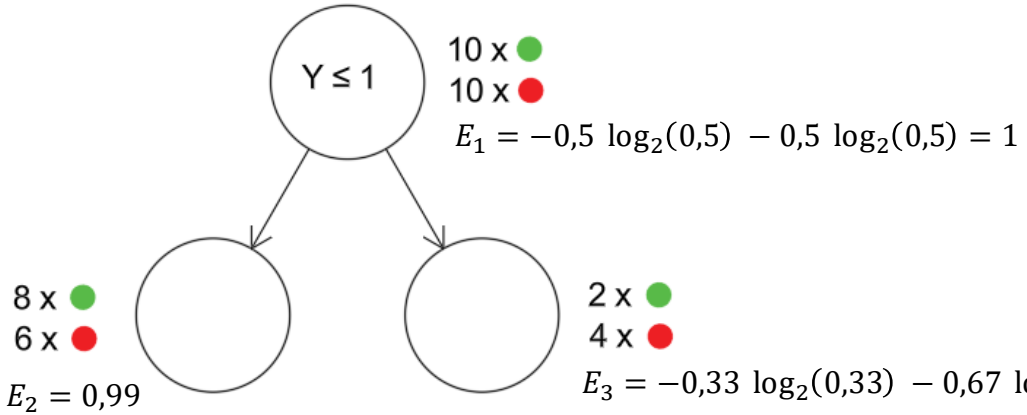


Wybór najlepszego podziału



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

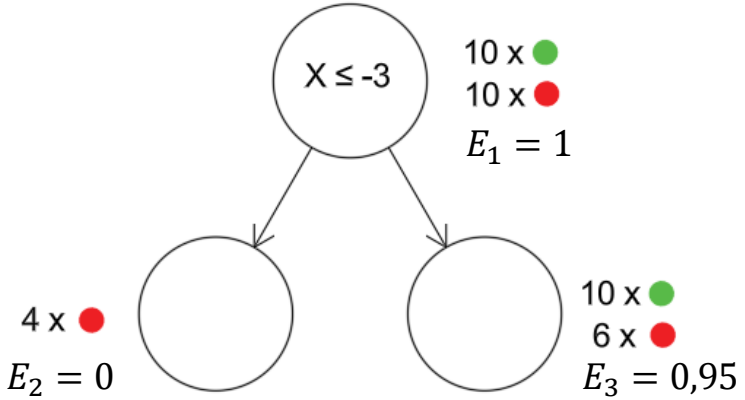
gdzie: n – liczba klas w zbiorze danych
p_i - prawdopodobieństwo wystąpienia klasy i w zbiorze danych



$E_1 = -0{,}5\ \log_2(0{,}5)\ - 0{,}5\ \log_2(0{,}5) = 1$

$E_2 = 0{,}99$

$E_3 = -0{,}33\ \log_2(0{,}33)\ - 0{,}67\ \log_2(0{,}5) = 0{,}91$



$E_1 = 1$

$E_2 = 0$

$E_3 = 0{,}95$

$$IG = E_{parent} - \sum w_i E_{child_i}$$

$$IG_1 = 1 - \frac{14}{20} \cdot 0{,}99 - \frac{6}{20} \cdot 0{,}91 = 0{,}034$$

$$IG_2 = 1 - \frac{4}{20} \cdot 0 - \frac{16}{20} \cdot 0{,}95 = 0{,}24$$

$$IG_2 > IG_1$$

Zalety drzewa decyzyjnego

ŁATWOŚĆ INTERPRETACJI

SZYBKOŚĆ DZIAŁANIA

ODPORNOŚĆ NA OBSERWACJE ODSTAJĄCE

WYMAGA NIEWIELKIEGO PRZYGOTOWANIA DANYCH

Wady drzewa decyzyjnego

Wysoka liczba rozgałęzień w złożonych drzewach

Wrażliwość na zmiany danych

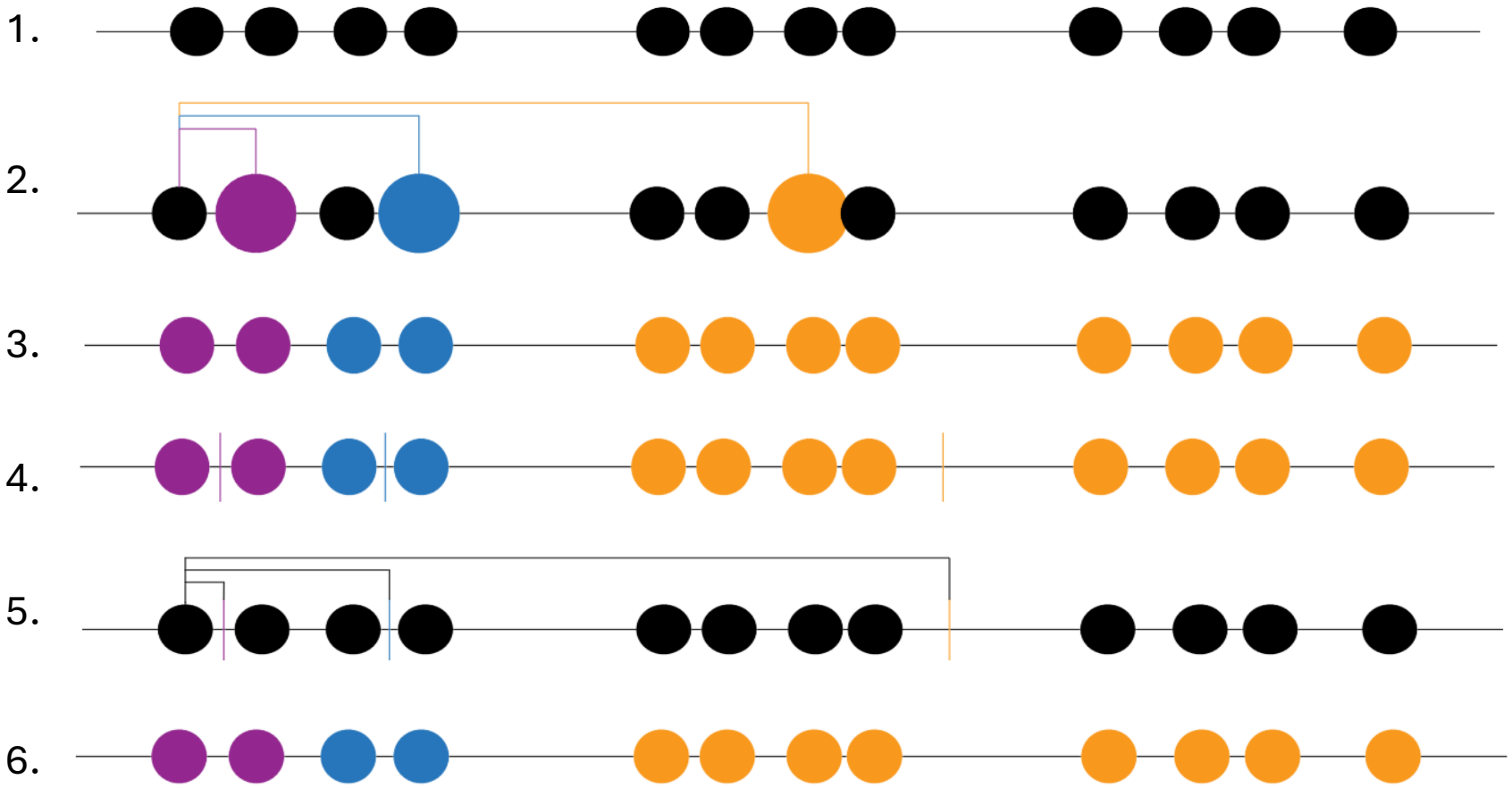
Przetrenowanie

KLASTERYZACJA K-ŚREDNICH

Cel

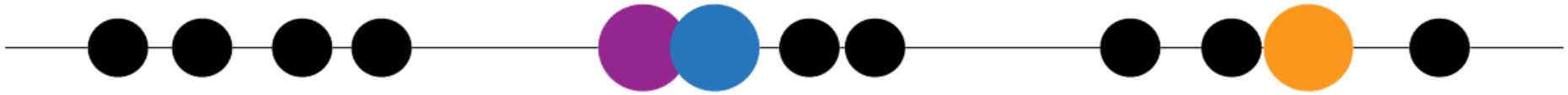
Klasteryzacja K-średnich jest jednym z najpopularniejszych i najprostszych algorytmów do grupowania danych. Jest szczególnie efektywna przy klasteryzacji dużych, liczbowych zbiorów danych o stosunkowo dobrze odseparowanych klastrach. Klasteryzacja k-średnich dąży do podziału zbioru danych na *k* klastrów wykorzystując do tego celu centroidę.

Zasada działania

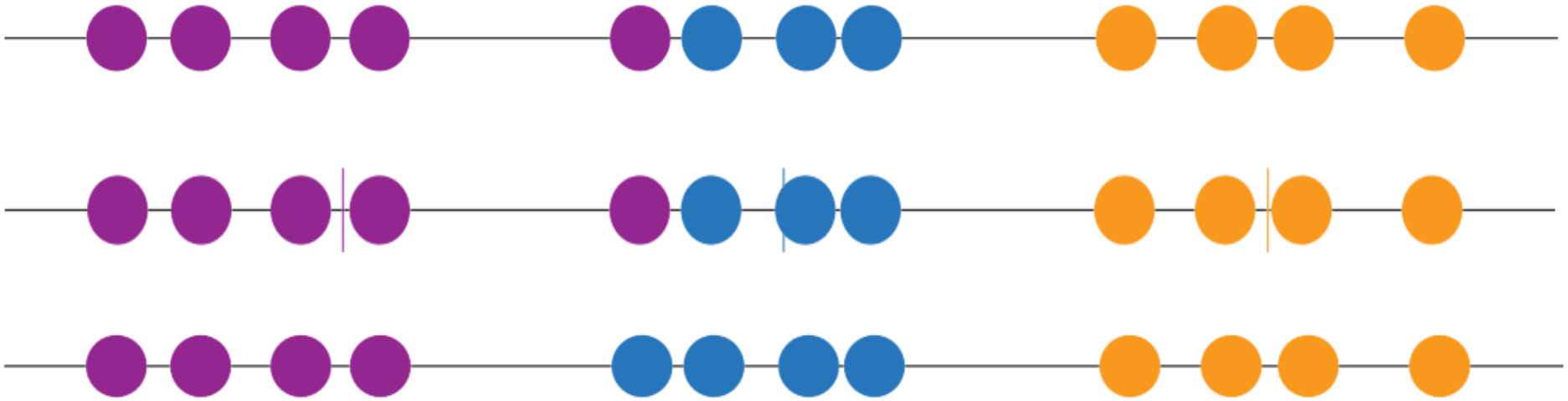


Zasada działania c.d.

Wybór nowych początkowych centroid:



Efekt:



Zalety oraz wady klasteryzacji k-średnich

Zalety:

- Prostota i szybkość działania
- Elastyczność
- Skalowalność

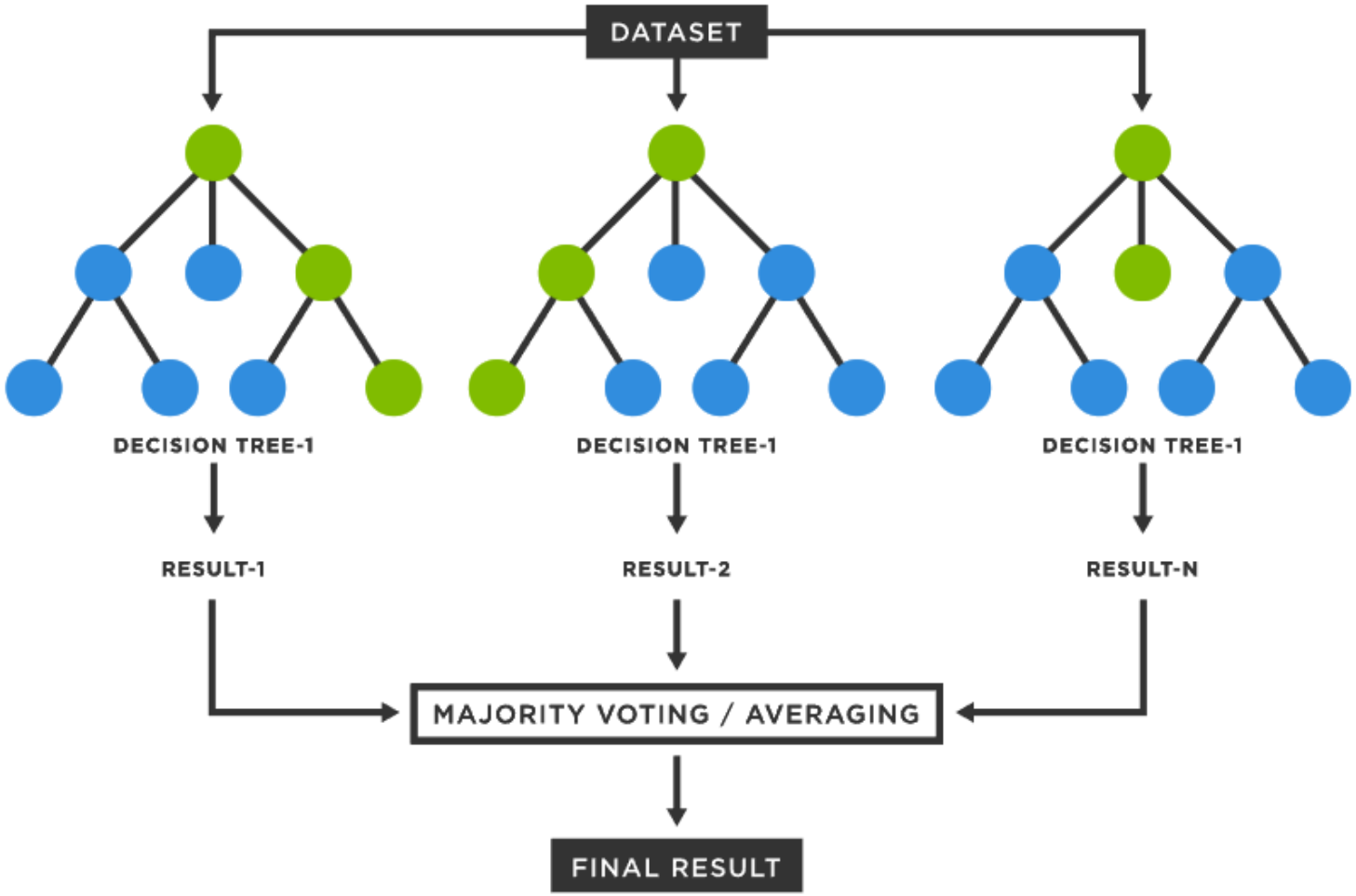
Wady:

- Wybór liczby klastrów
- Wrażliwe na dobór punktów startowych
- Używa jedynie zmiennych numerycznych

Las Losowy

Las losowy polega na tworzeniu zbioru drzew klasyfikacyjnych, uzyskiwanych dzięki losowemu wyborowi grupy zmiennych ze zbioru zmiennych oraz metodzie *bootstrap*, bazującej na losowym wybieraniu części próbek ze zbioru danych, aby stworzyć wiele zestawów treningowych dla każdego drzewa w modelu. Na każdym etapie budowy drzewa wybierane są najbardziej rozróżniające zmienne, aż do momentu, gdy na końcach drzewa znajdują się czyste klasy.

Struktura Lasu Losowego



Las Losowy

Zalety:

- Wysoka dokładność i odporność na przeuczenie
- Radzenie sobie z dużą liczbą cech
- Odporność na szum
- Niewielkie wymagania co do przygotowania danych
- Znaczenie cech

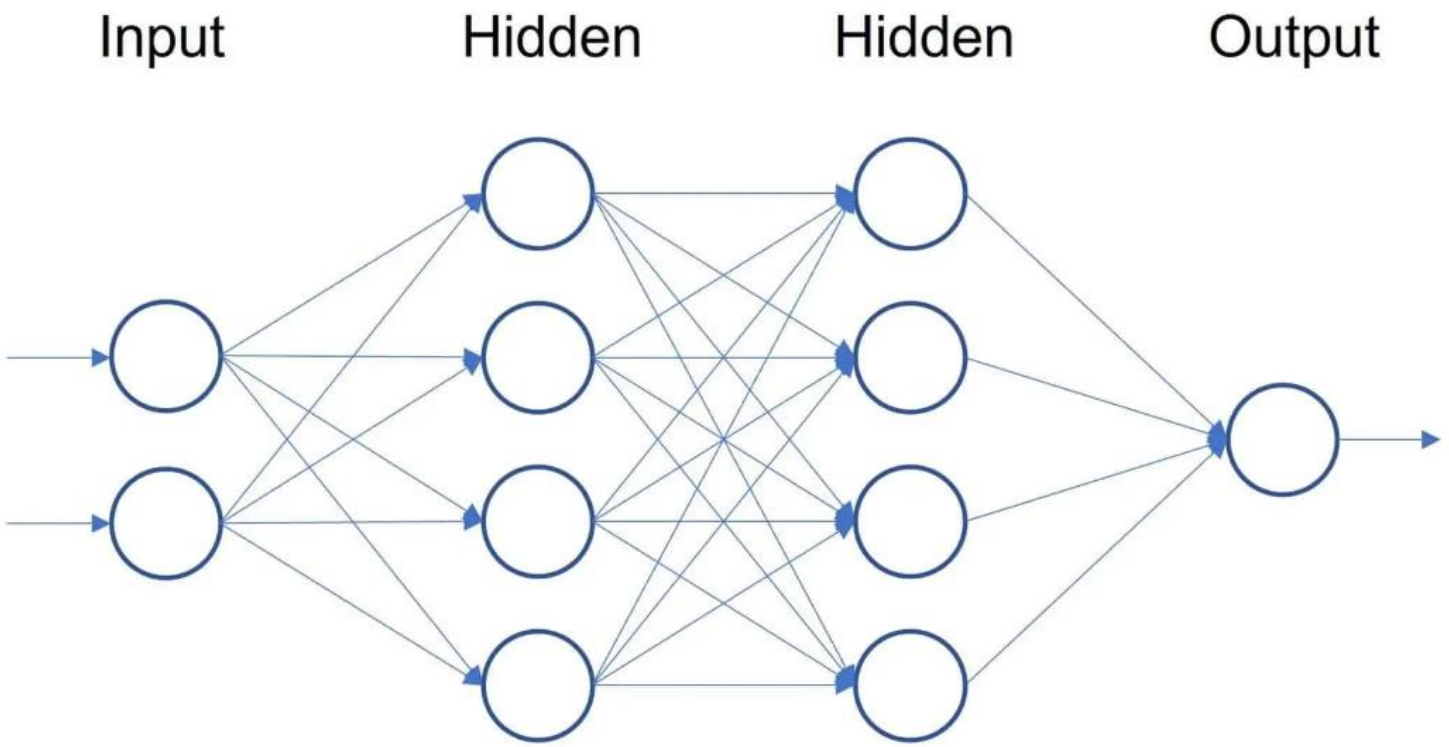
Wady:

- Duże zasoby obliczeniowe
- Brak interpretowalności
- Mniejsza efektywność dla danych o wysokiej liczbie wymiarów

Sieci Neuronowe

Sieci Neuronowe to modele matematyczne inspirowane sposobem, w jaki funkcjonuje ludzki mózg. Składają się one z wielu połączonych ze sobą jednostek zwanych neuronami, które przetwarzają dane i przekazują wyniki do kolejnych neuronów, tworząc strukturę podobną do neuronów w ludzkim mózgu.

Struktura Sieci Neuronowej



Sieci Neuronowe

Zalety:

- Rozpoznanie złożonych wzorców
- Wysoka wydajność przy dużej ilości danych
- Uczenie bez nadzoru
- Automatyczne tworzenie cech
- Wszechstronność i elastyczność

Wady:

- Wysokie wymagania obliczeniowe
- Wymaganie dużych zbiorów danych
- Brak interpretowalności
- Ryzyko przeuczenia
- Złożoność i trudność w dostrajaniu
- Wysoki koszt energetyczny

Bibliografia

[3-5] https://learn.microsoft.com/pl-pl/training/modules/understand-regression-machine-learning/

[6] https://www.easiio.com/pl/easiio-linear-regression-in-machine-learning/

[7-8] https://predictivesolutions.pl/regresja-logistyczna

[9] https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

[9-10] https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

[12,13,17,18] Rokach, L., & Maimon.: *Data Mining with Decision Trees: Theory and Applications*. World Scientific (2008)

[20,24] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y.: *An Efficient k-Means Clustering Algorithm: Analysis and Implementation*. (2002)

[27] Conesa, A., & Hernández, R. (2014). *Omics Data Integration in Systems Biology. Applications of Advanced Omics Technologies: From Genes to Metabolites, 441–459*. doi:10.1016/b978-0-444-62650-9.00016-6

[28] https://medium.com/@dishantkharkar9/about-random-forest-algorithms-62163357db25

[29] https://www.geeksforgeeks.org/what-are-the-advantages-and-disadvantages-of-random-forest/

[30,32] Tu, J. V. (1996). *Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology, 49(11), 1225–1231*. doi:10.1016/s0895-4356(96)00002-9

[31] https://www.cienciasinseso.com/en/neural-networks/