dr inż. Michał Malinowski

https://www.drmalinowski.edu.pl

Temat 5 .

Jak uczą się maszyny? Przegląd kluczowych algorytmów AI

Wprowadzenie do Sztucznej Inteligencji



- Algorytmy uczenia maszynowego
- Hiperparametry
- Regresja liniowa i Logistyczna
- Drzewa Decyzyjne
- SVM – Maszyny Wektorów Nośnych
- Las Losowy
- K-najbliższych sąsiadów
- K-średnich
- Naiwny Klasyfikator Bayesa
- Sztuczne Sieci Neuronowe
- Algorytm Apriori
- Algorytmy Genetyczne
- Algorytmy Grafowe

Czas stracony na szukanie (*Findability problem*)



Pracownicy wiedzy spędzają **do 30% czasu pracy** na poszukiwaniu informacji niezbędnych do wykonania zadań.
IDC – The High Cost of Not Finding Information (2012)

50% firm przyznaje, że ich pracownicy mają trudności ze znalezieniem właściwej wersji dokumentów.
AIIM – State of Information Management (2016)

Pracownik spędza średnio **1,8 godziny dziennie (22,5%)** na szukaniu i gromadzeniu potrzebnych informacji.
McKinsey Global Institute – The social economy (2012)

Nawet **40% czasu pracy** może być przeznaczane na odnalezienie właściwych danych w rozproszonych systemach.
Gartner (2013)

Zjawisko zatrzymywania wiedzy (*Knowledge hoarding*)

Strach przed **utratą kontroli i władzy** jest istotnym czynnikiem ograniczającym dzielenie się wiedzą.
Renzl, B. – "Trust in management and knowledge sharing: The mediating effects of fear and knowledge documentation„ (2008)

Wrogie nastawienie do dzielenia się wiedzą wynika z postrzegania informacji jako **waluty władzy organizacyjnej**.
Husted, K. & Michailova, S. – "Diagnosing and fighting knowledge-sharing hostility„ (2002)

Wiedza jest często traktowana jako **zasób strategiczny**, co powoduje, że osoby ją posiadające **nie chcą się nią dzielić bezpośrednio**.
Nonaka, I., Takeuchi, H. – "The Knowledge-Creating Company„ (1995)



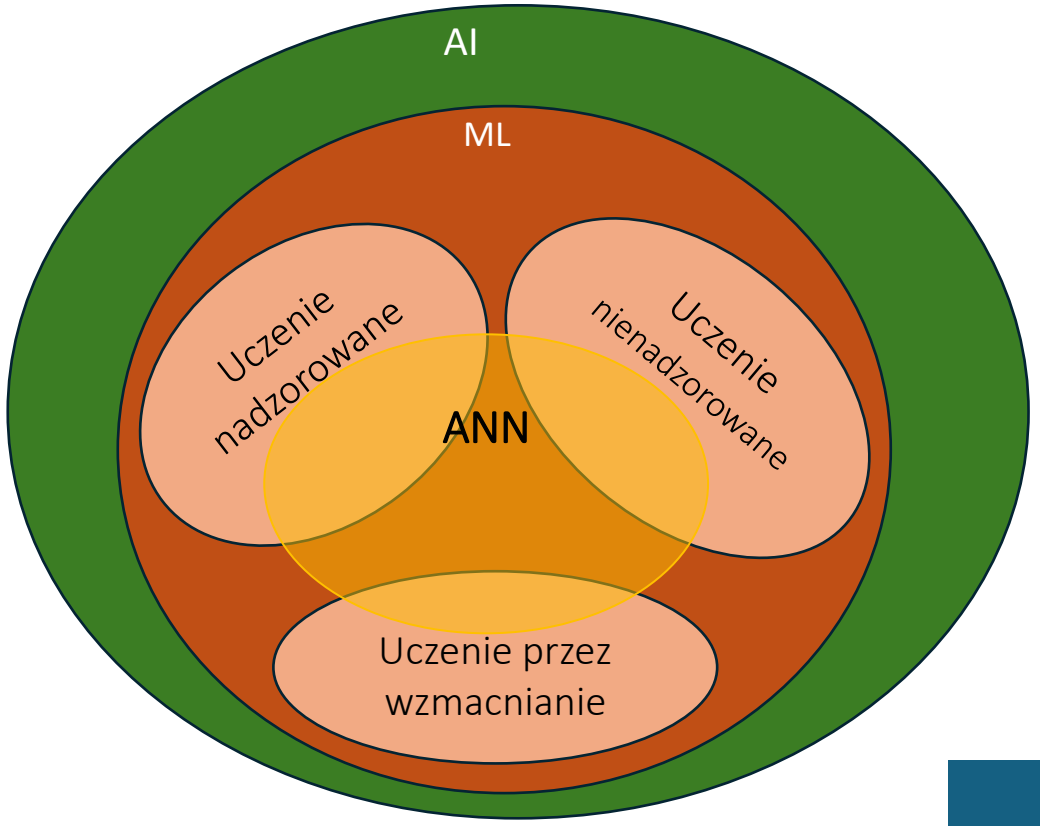
Czy u Ciebie w organizacji zachodzi to zjawisko ?

Algorytmy uczenia maszynowego

Algorytmy uczenia maszynowego (Machine Learning Algorithms)

to specjalne metody matematyczne i statystyczne, które umożliwiają komputerom
analizowanie danych, wykrywanie wzorców i podejmowanie decyzji bez konieczności jawnego programowania.

Uczenie maszynowe (ML) jest częścią **sztucznej inteligencji (AI)** i pozwala modelom poprawiać swoje działanie na podstawie doświadczenia (czyli nowych danych), zamiast korzystać ze sztywnych reguł określonych przez programistę.



ANN - Sztuczne Sieci Neuronowe

Hiperparametry

Hiperparametry

to parametry zewnętrzne modelu uczenia maszynowego, które są ustawiane przed rozpoczęciem procesu uczenia i nie są bezpośrednio optymalizowane podczas tego procesu. Różnią się one od parametrów modelu, które są nauczane podczas procesu uczenia się z danych.

Liczba warstw i neuronów w sieci neuronowej:

Określa, ile warstw (głębokości) i ile neuronów (szerokości) ma sieć neuronowa.

Szybkość uczenia:

Określa, jak duże kroki są wykonywane podczas aktualizacji wag modelu.

Liczba epok:

Określa, ile razy cały zbiór danych treningowych przechodzi przez model.

Rozmiar partii (batch size):

Określa liczbę próbek danych przetwarzanych przed aktualizacją wag modelu.

Parametr regularyzacji:

Pomaga w uniknięciu przeuczenia modelu poprzez dodanie kary za złożoność modelu.

Hiperparametry

Przykłady

Hiperparametr	Wartość niska	Wartość wysoka
Liczba epok	Model niedotrenowany, za mało epok do nauki wzorców	Model przeuczony – uczy się zbyt długo, zapamiętuje dane zamiast generalizować
Szybkość uczenia	Model uczy się bardzo wolno, może utknąć w minimum lokalnym	Zbyt agresywna nauka powoduje niestabilność i słabsze wyniki
Rozmiar partii (batch size)	Niższa stabilność optymalizacji, większa różnorodność aktualizacji wag	Szybszy trening, ale ryzyko przeoczenia ważnych cech danych
Liczba warstw i neuronów	Model zbyt prosty, nie uczy się wystarczająco skomplikowanych zależności	Model zbyt złożony, może łatwo dopasować się do szumu
Parametr regularyzacji	Model może nadmiernie dopasowywać się do danych treningowych	Model może stać się zbyt ogólny i nie rozpoznać subtelnych wzorców

Regresja liniowa

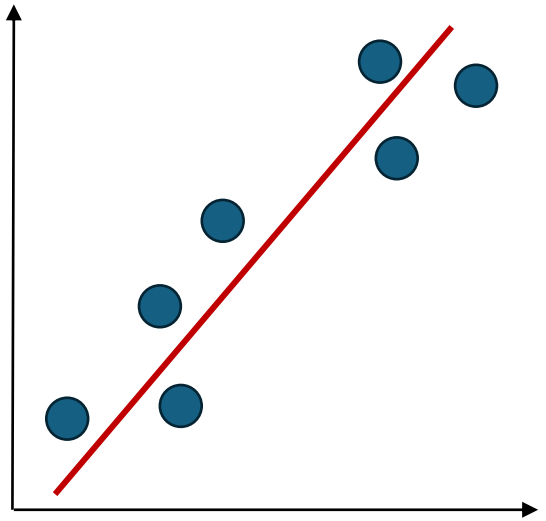
Regresja liniowa (Linear Regression)

to **podstawowy algorytm uczenia maszynowego**, który **szuka prostej linii najlepiej dopasowującej się do danych**.

Jest stosowany do **przewidywania wartości liczbowych (regresji)** na podstawie jednej lub wielu zmiennych niezależnych.

Zastosowanie

🏠 **Przewidywanie cen domów** na podstawie cech takich jak metraż, liczba pokoi itp.
📊 **Analiza trendów rynkowych** – prognozowanie sprzedaży w zależności od czynników ekonomicznych.
⚕️ **Modele predykcyjne w medycynie** – szacowanie ryzyka chorób na podstawie danych pacjenta.



Zadanie

Zwróć uwagę na fragment kodu budujący model AI



Treść

W Google Colab https://colab.research.google.com/ stwórz prosty system uczenia maszynowego bazujący na algorytmie regresji liniowej.

1. Wygeneruj losowy zbiór danych (100 pozycji) o cenach domów w zależności od metrażu i lokalizacji (miasta).
2. Wytrenuj model regresji liniowej na tym zbiorze danych.
3. Wygeneruj 5 losowych domów (metraż i lokalizacja).
4. Przewidź dla tych 5 losowych domów ceny za pomocą wytrenowanego modelu.

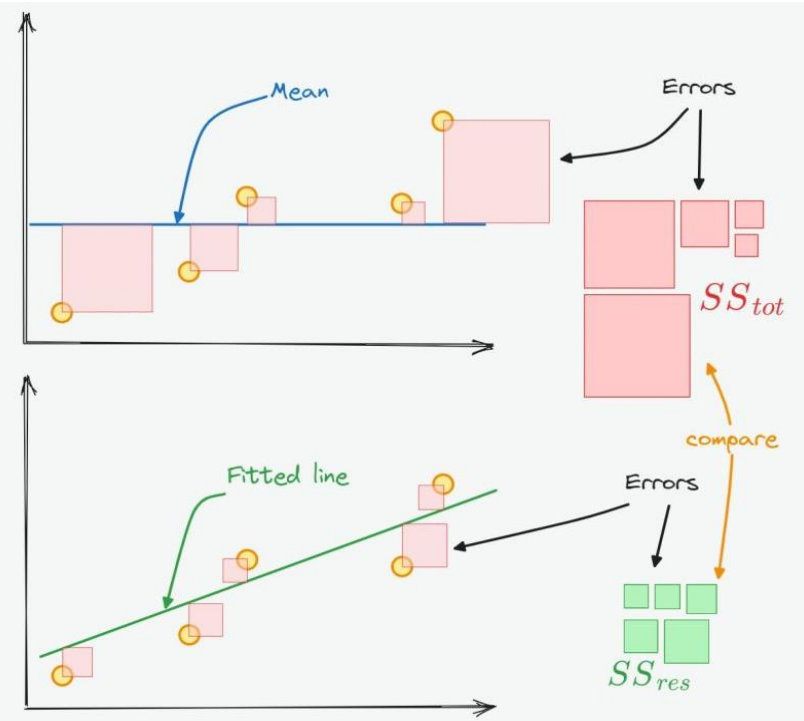
https://colab.research.google.com/drive/1P84TUEoKoBmcC7rDYGcz5f2JAhPVUk4E?usp=sharing

Miary regresji liniowej

R^2 (R-Squared, współczynnik determinacji)

to miara statystyczna, która określa, jak dobrze model regresyjny dopasowuje się do danych.
Innymi słowy, pokazuje, jaka część zmienności zmiennej zależnej (np. ceny domu) jest wyjaśniona przez zmienne niezależne (np. metraż, lokalizacja).

Interpretacja wartości R^2:
🔴 $R^2 < 0$ → Model jest gorszy niż losowe zgadywanie
🔴 $R^2 = 0$ → Model nie wyjaśnia w ogóle zmienności danych, prognozy są losowe.
🟢 $0 < R^2 < 1$ → Model dobrze dopasowany – zmienne niezależne wyjaśniają część zmienności zmiennej zależnej, ale istnieje pewien poziom błędu. Im wyższa wartość, tym lepsze dopasowanie.
🟢 $R^2 = 1$ → Model idealnie przewiduje wartości, wszystkie punkty leżą na linii regresji.



$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

$$SS_{res} = \sum_i (y_i - f_i)^2$$

Miary regresji liniowej

Mean Squared Error (MSE)

to jedna z najczęściej stosowanych metryk do oceny jakości modeli regresyjnych.

Mierzy średni kwadrat różnic między wartościami rzeczywistymi a przewidywanymi przez model.

Interpretacja:

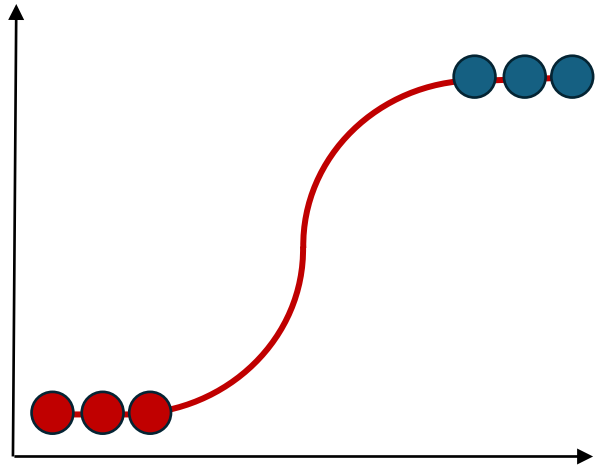
- MSE = 0 → Model idealnie przewiduje wartości.
- Niższe wartości MSE oznaczają lepsze dopasowanie modelu.
- Wyższe wartości MSE wskazują, że model generuje większe błędy.

Regresja Logistyczna

Regresja Logistyczna (Binarna) (Logistic Regression)

to **model klasyfikacyjny**, który przewiduje, czy dana obserwacja **należy do jednej z dwóch kategorii** (np. **„tak”-„nie”**, **„0”-„1”**).

Zwraca prawdopodobieństwo przynależności do jednej z klas, wykorzystując funkcję **sigmoidalną (logistyczną)** do przekształcenia wyników.



Zastosowanie

🛍️ **Przewidywanie, czy klient dokona zakupu** na podstawie jego historii zakupów.
📧 **Filtrowanie spamu** – klasyfikacja e-maili jako spam/niespam.
🏥 **Diagnozowanie chorób** – przewidywanie, czy pacjent ma określoną chorobę (np. cukrzycę) na podstawie wyników badań.
⚖️ **Modele decyzyjne w bankowości** – klasyfikacja wniosków kredytowych jako „zaakceptowany” lub „odrzucony”.

Zadanie

Zwróć uwagę na fragment kodu budujący model AI



Treść

W Google Colab https://colab.research.google.com/ stwórz prosty system uczenia maszynowego bazujący na algorytmie regresji logistycznej.

1. Wygeneruj losowy zbiór danych (100 pozycji) o klientach starających się o kredyt. Dane powinny zawierać: wiek klienta, dochód miesięczny, decyzja kredytowa (1 = przyznano kredyt, 0 = odmowa)
2. Wytrenuj model regresji logistycznej na tym zbiorze danych.
3. Wygeneruj 5 losowych klientów w postaci tablicy bez indeksów (wiek, dochód).
4. Przewidź dla tych 5 klientów decyzję kredytową (przyznano[1]/odmowa[0]) za pomocą wytrenowanego modelu.

https://colab.research.google.com/drive/1b-w_TGgWkxH3fAY8o8YkvqkOz95auHUF?usp=sharing

Miary regresji logistycznej

Accuracy (dokładność)

to miara oceniająca skuteczność modelu klasyfikacyjnego, w tym regresji logistycznej.

Jest obliczana jako stosunek poprawnie przewidzianych etykiet do wszystkich próbek w zbiorze testowym.

Interpretacja

- Accuracy = 1.0 (100%) → Model przewidział wszystkie klasy poprawnie.
- Accuracy = 0.8 (80%) → Model poprawnie przewidział 80% przypadków.
- Accuracy = 0.5 (50%) → Model przewiduje losowo, co oznacza słabe

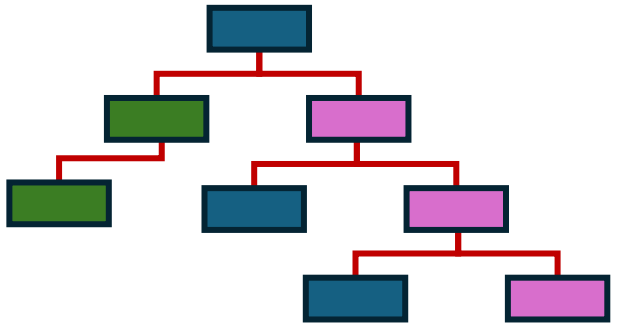
Drzewa Decyzyjne

Drzewa Decyzyjne (Decision Tree)

to model, który **podejmuje decyzje poprzez podział danych na mniejsze grupy**, tworząc **ścieżkę wyborów** opartą na wartościach cech.

Może być używany zarówno do **klasyfikacji**, jak i **regresji**.

Zastosowanie



📱 **Klasyfikacja telefonów według półki cenowej** – Budżetowy / Średni / Premium na podstawie ceny i specyfikacji.
🛍️ **Segmentacja klientów** – Nowy / Okazjonalny / Lojalny na podstawie historii zakupów.
🏥 **Diagnoza choroby** – Grypa / Przeziębienie / COVID-19 na podstawie objawów.
🚗 **Klasyfikacja pojazdów** – Samochód / Motocykl / Ciężarówka na podstawie cech technicznych.

Zadanie

Zwróć uwagę na fragment kodu budujący model AI



Treść

W Google Colab https://colab.research.google.com/ stwórz prosty system uczenia maszynowego bazujący na drzewie decyzyjnym.

1. Wygeneruj losowy zbiór danych (100 pozycji) o pacjentach. Dane powinny zawierać: Wiek pacjenta, BMI (wskaźnik masy ciała), Poziom cukru we krwi (mg/dL), Ryzyko choroby (1 = wysokie, 0 = niskie)
2. Wytrenuj model drzewa decyzyjnego na tym zbiorze danych.
3. Wygeneruj 5 losowych pacjentów dla przetestowania modelu (wiek, BMI, poziom cukru).
4. Przewidź dla tych 5 pacjentów ryzyko choroby (wysokie [1] / niskie [0]) za pomocą wytrenowanego modelu.
5. Zwizualizuj strukturę drzewa decyzyjnego.

https://colab.research.google.com/drive/1sDB_zkWVCOU1Svbx6hFpL2p6nb-vJmoe?usp=sharing

SVM – Maszyny Wektorów Nośnych

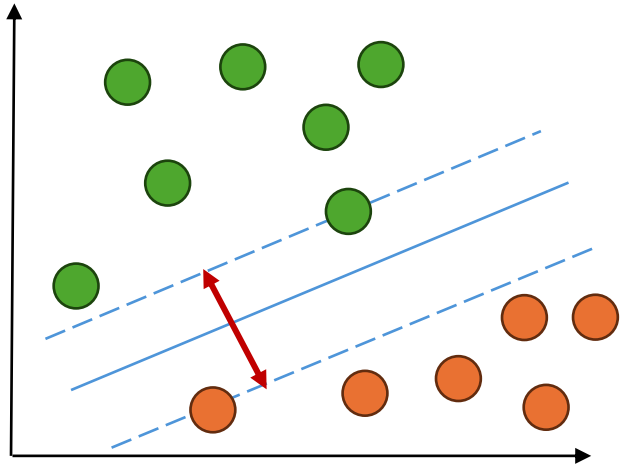
SVM – Maszyny Wektorów Nośnych (Support Vector Machines)

to **algorytm klasyfikacji**, który znajduje **optymalną linię (dla 2D) lub płaszczyznę (dla 3D+)**, najlepiej **oddzielającą dane** należące do różnych klas.

Model ten **maksymalizuje margines** między klasami, co pomaga w lepszym rozpoznawaniu wzorców.

Zastosowanie

📷 **Klasyfikacja zdjęć zwierząt** – rozpoznawanie, czy na obrazie jest kot czy pies.
📝 **Rozpoznawanie pisma** – klasyfikacja ręcznie napisanych cyfr i liter.
🧬 **Analiza genomu** – klasyfikowanie sekwencji DNA.
📡 **Wykrywanie anomalii w cyberbezpieczeństwie** – identyfikowanie podejrzanych aktywności w sieci.



Las Losowy

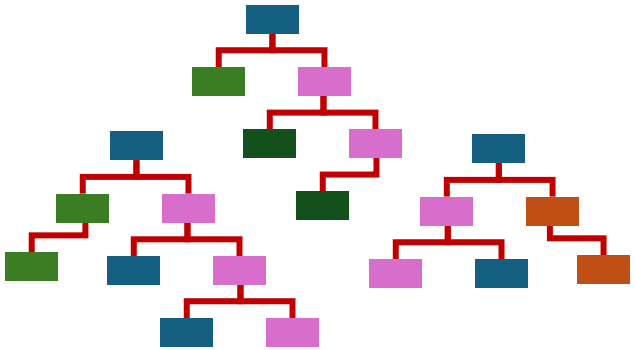
Las Losowy (Random Forest)

to **model**, który składa się z **wielu drzew decyzyjnych**, współpracujących w celu **zwiększenia dokładności prognoz** i ograniczenia ryzyka przeuczenia.

Każde drzewo analizuje losową część danych i głosuje na końcowy wynik w ramach **klasyfikacji** lub **regresji**.

Zastosowanie

❤️ **Przewidywanie, czy pacjent ma chorobę serca** na podstawie danych medycznych.
🏢 **Ocena ryzyka kredytowego** – prognozowanie zdolności kredytowej klientów.
🛒 **Rekomendacje zakupowe** – analiza wzorców zachowań użytkowników.
🌾 **Przewidywanie plonów w rolnictwie** – analiza czynników wpływających na wielkość zbiorów.



K-najbliższych sąsiadów

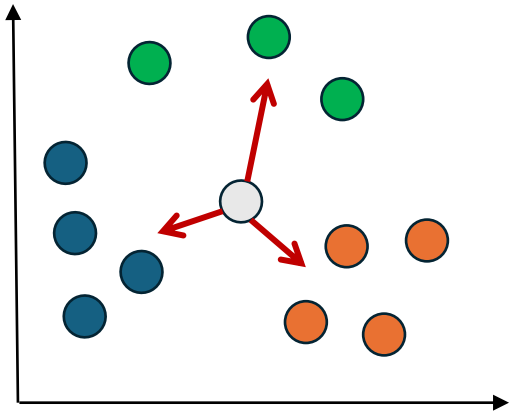
K-najbliższych sąsiadów (k-Nearest Neighbors, KNN)

algorytm przewiduje **kategorię nowego punktu**, analizując **k najbliższych sąsiadów** z danych treningowych.

W zależności od większościowej klasy sąsiadów, nowa próbka zostaje **zaklasyfikowana** do odpowiedniej grupy.

Zastosowanie

🎬 **Rekomendacje filmów** – przewidywanie, czy nowy film spodoba się użytkownikowi, na podstawie ocen innych osób.
📷 **Rozpoznawanie obrazów** – klasyfikowanie zdjęć na podstawie wzorców kolorów i kształtów.
🏢 **Wykrywanie oszustw bankowych** – analiza transakcji pod kątem podobieństwa do znanych przypadków oszustw.



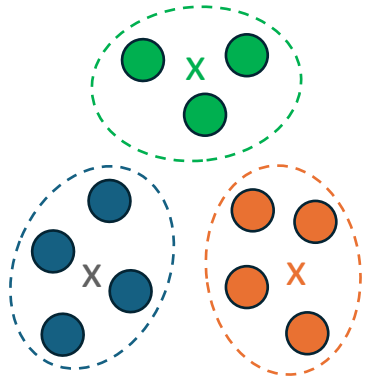
K-średnich

K-średnich (K-means Clustering)

to algorytm **grupowania (klasteryzacji)**, który dzieli dane na **k klastrów**. Każdy punkt zostaje przypisany do najbliższego **centrum klastra (centroidu)**, a algorytm iteracyjnie aktualizuje pozycje centroidów, aby lepiej dopasować podział.

Zastosowanie

🛒 **Segmentacja klientów** – podział klientów sklepu na grupy, np. **„często kupujący”, „nowi klienci”**.
🎨 **Kompresja obrazów** – redukcja kolorów poprzez grupowanie pikseli o podobnych barwach.
🌍 **Analiza geograficzna** – grupowanie lokalizacji na podstawie podobieństw (np. miasta o podobnych wzorcach pogodowych).



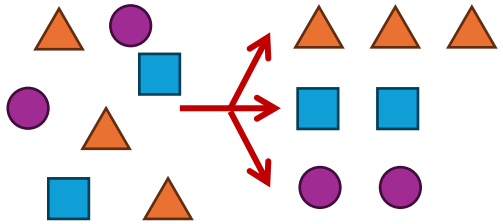
Naiwny Klasyfikator Bayesa

Naiwny Klasyfikator Bayesa (Naive Bayes)

to **klasyfikator probabilistyczny**, który opiera się na **twierdzeniu Bayesa**.

Zakłada, że **wszystkie cechy są niezależne**, co upraszcza obliczenia i pozwala na szybką klasyfikację.

Zastosowanie



💬 **Analiza opinii** – klasyfikacja komentarzy jako **„pozytywne” lub „negatywne”**.
📧 **Filtrowanie spamu** – wykrywanie niechcianych e-maili na podstawie ich treści.
🩺 **Diagnoza medyczna** – przewidywanie chorób na podstawie objawów pacjenta.
🔍 **Rozpoznawanie kategorii tekstów** – klasyfikacja dokumentów według tematyki (np. „sport”, „nauka”, „polityka”).

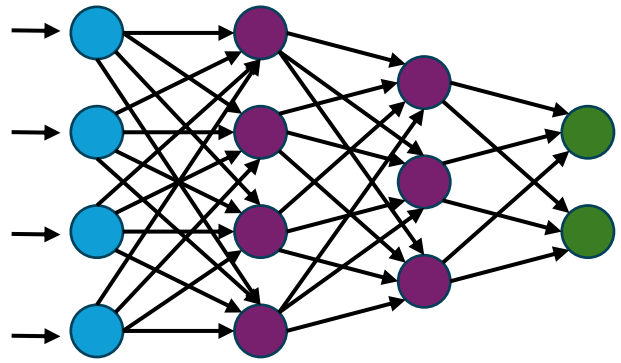
Sztuczne Sieci Neuronowe

Sztuczne Sieci Neuronowe (Artificial Neural Networks, ANN)

to **zaawansowane modele inspirowane działaniem ludzkiego mózgu**. Składają się z **warstw neuronów**, które uczą się skomplikowanych zależności w danych. Są szeroko stosowane do **rozpoznawania obrazów, analizy języka i prognozowania danych**..

Zastosowanie

📸 **Rozpoznawanie twarzy na zdjęciach i filmach** – np. odblokowywanie telefonu za pomocą Face ID.
📝 **Przetwarzanie języka naturalnego (NLP)** – tłumaczenia, chatboty, analiza sentymentu.
🩺 **Diagnostyka medyczna** – analiza zdjęć rentgenowskich i MRI.
🎮 **Sztuczna inteligencja w grach** – sterowanie postaciami NPC.



Algorytm Apriori

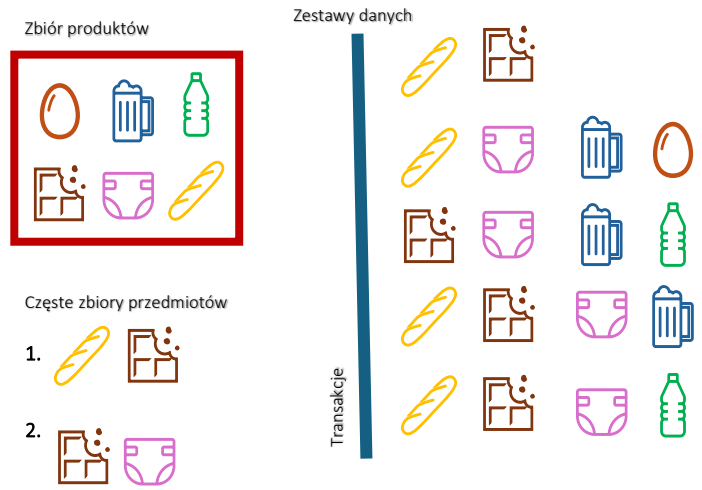
Algorytm Apriori (Apriori Algorithm)

to **metoda eksploracji danych**, która umożliwia **odkrywanie ukrytych wzorców i zależności** w dużych zbiorach danych.

Pozwala na **identyfikację reguł asocjacyjnych**, które mogą wspierać podejmowanie decyzji biznesowych.

Zastosowanie

🛍️ **Analiza koszykowa** – wykrywanie wzorców zakupowych klientów, np. osoby kupujące chleb często kupują również masło.
🤖 **Systemy rekomendacji** – sugerowanie produktów na podstawie wcześniejszych wyborów użytkowników (np. "Klienci, którzy kupili ten produkt, kupili także...").
📄 **Analiza treści** – identyfikowanie powiązań między tematami dokumentów i kategoryzacja tekstów.



Algorytmy Genetyczne

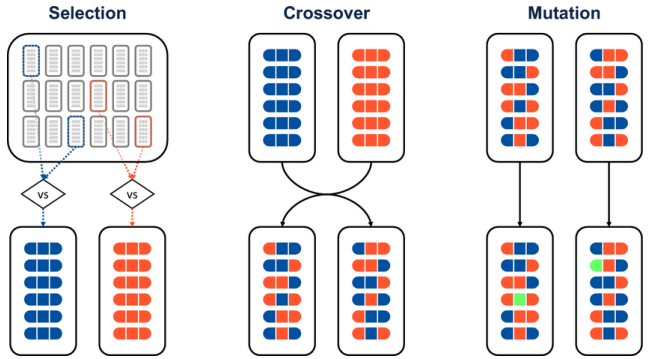
Algorytmy Genetyczne (Genetic Algorithms, GA)

to **heurystyczna metoda optymalizacji**, inspirowana **procesami ewolucyjnymi** w naturze.

Wykorzystują **selekcję, mutację i krzyżowanie**, aby znajdować najlepsze rozwiązania problemów.

Zastosowanie

🛠 **Optymalizacja inżynieryjna** – projektowanie efektywnych struktur, np. **optymalne skrzydło samolotu**.
📈 **Finanse i handel algorytmiczny** – znajdowanie **najlepszych strategii inwestycyjnych**.
🧬 **Bioinformatyka** – analiza **sekwencji DNA** i dobór najlepszych kombinacji genów.
📊 **Rozwiązywanie problemów NP-trudnych** – takich jak **problem komiwojażera** (najkrótsza trasa dla podróżującego sprzedawcy).



Algorytmy Genetyczne

Algorytmy Genetyczne (Genetic Algorithms, GA)

🔷 Dlaczego GA można zaliczyć do ML?

✓ **Są metodą optymalizacji inspirowaną ewolucją biologiczną**
używają selekcji, mutacji i krzyżowania do znajdowania najlepszych rozwiązań.

✓ **Uczą się na podstawie iteracyjnego ulepszania populacji rozwiązań**
podobnie jak inne algorytmy ML, dążą do optymalizacji funkcji celu.

✓ **Stosowane w połączeniu z innymi algorytmami ML**
np. do strojenia hiperparametrów sieci neuronowych.

🔻 Dlaczego GA nie są klasycznym algorytmem ML?

❌ **Nie uczą się bezpośrednio z danych**
w przeciwieństwie do klasycznych algorytmów ML, które wykrywają wzorce w danych wejściowych.

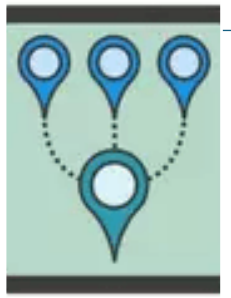
❌ **Nie wykorzystują statystycznych metod uczenia**
zamiast tego opierają się na heurystykach inspirowanych ewolucją.

❌ **Często używane w optymalizacji, a nie w predykcji**
służą np. do znajdowania najlepszych strategii w problemach nieliniowych.

Algorytmy uczenia maszynowego

	Klasyfikacja	Regresja	Klasteryzacja	Asocjacja	Redukcja wymiarowości	Generowanie
Regresja liniowa (OLS)		tak				
Regresja logistyczna (Binarna)	tak	tak				
Drzewa decyzyjne	tak					
Las losowy	tak					
SVM (Maszyny wektorów nośnych)	tak					
K-najbliższych sąsiadów	tak					
K-średnich			tak			
PCA (Analiza składowych głównych)					tak	
Sztuczne sieci neuronowe	tak	tak	tak	tak	tak	tak
Naiwny klasyfikator Bayesa	tak					
Algorytm Apriori				tak		
Algorytmy genetyczne						Optymalizacja

Algorytmy grafowe



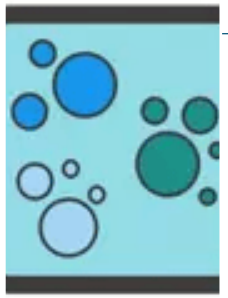
Algorytmy wyszukiwania ścieżek

Wyszukują optymalne ścieżki pomiędzy węzłami dla określonych kryteriów



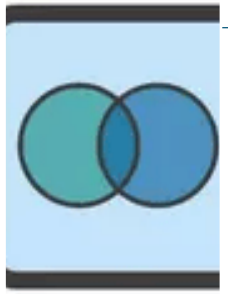
Algorytmy centralności

Pozwalają ocenić, jakie węzły w sieci są najważniejsze, czyli mogą mieć największy wpływ



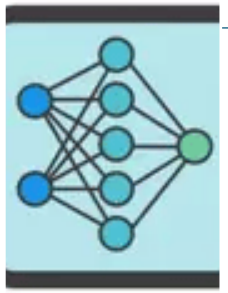
Algorytmy wspólnoty

Identifikują skupiska węzłów, które są gęściej połączone między sobą niż z resztą sieci



Algorytmy podobieństwa

Obliczają stopień podobieństwa między elementami



Algorytmy detekcji anomalii

Wykorzystywane są do identyfikowania nieoczekiwanych wzorców lub obserwacji w danych

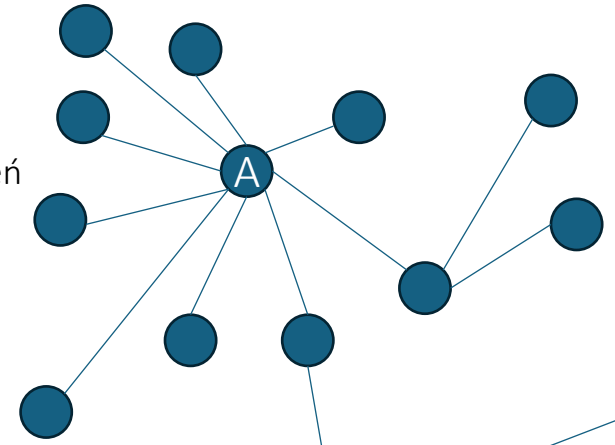


Algorytmy predykcji relacji

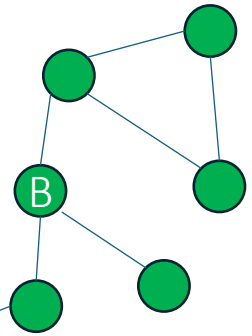
Prognozują czy dwa węzły w grafie mogą w przyszłości stworzyć połączenie

Algorytmy centralności

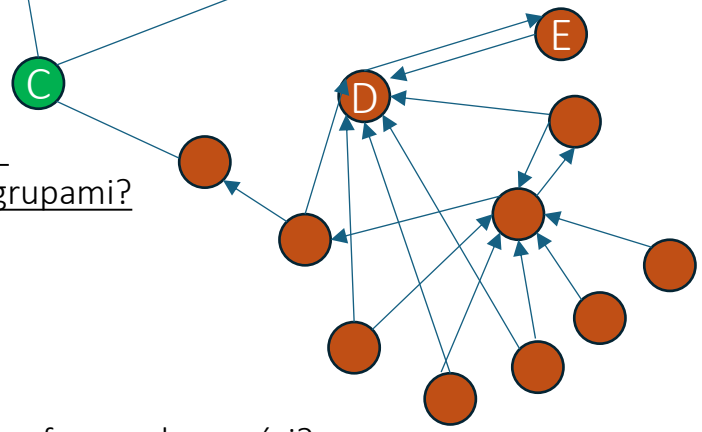
Degree Centrality
(Centralność stopnia)
Liczba połączeń?
"A" ma najwyższy stopień



Closeness Centrality
(Centralność pośrednictwa)
Który węzeł może najłatwiej dotrzeć do wszystkich innych węzłów w grafie?
"B" jest najbliższy z najmniejszą liczbą połączeń



Betweenness Centrality
(Centralność bliskości)
Który węzeł ma największą kontrolę nad przepływem między węzłami i grupami?
"C" jest mostem



PageRanking
Który węzeł jest najważniejszy?
"D" jest na czele na podstawie liczby i wagi linków przychodzących
"E" jest następny, ze względu na wpływ linku "D"

Louvain Modularity
(Modułowość Louvain)
Jak efektywnie podzielić graf na społeczności?



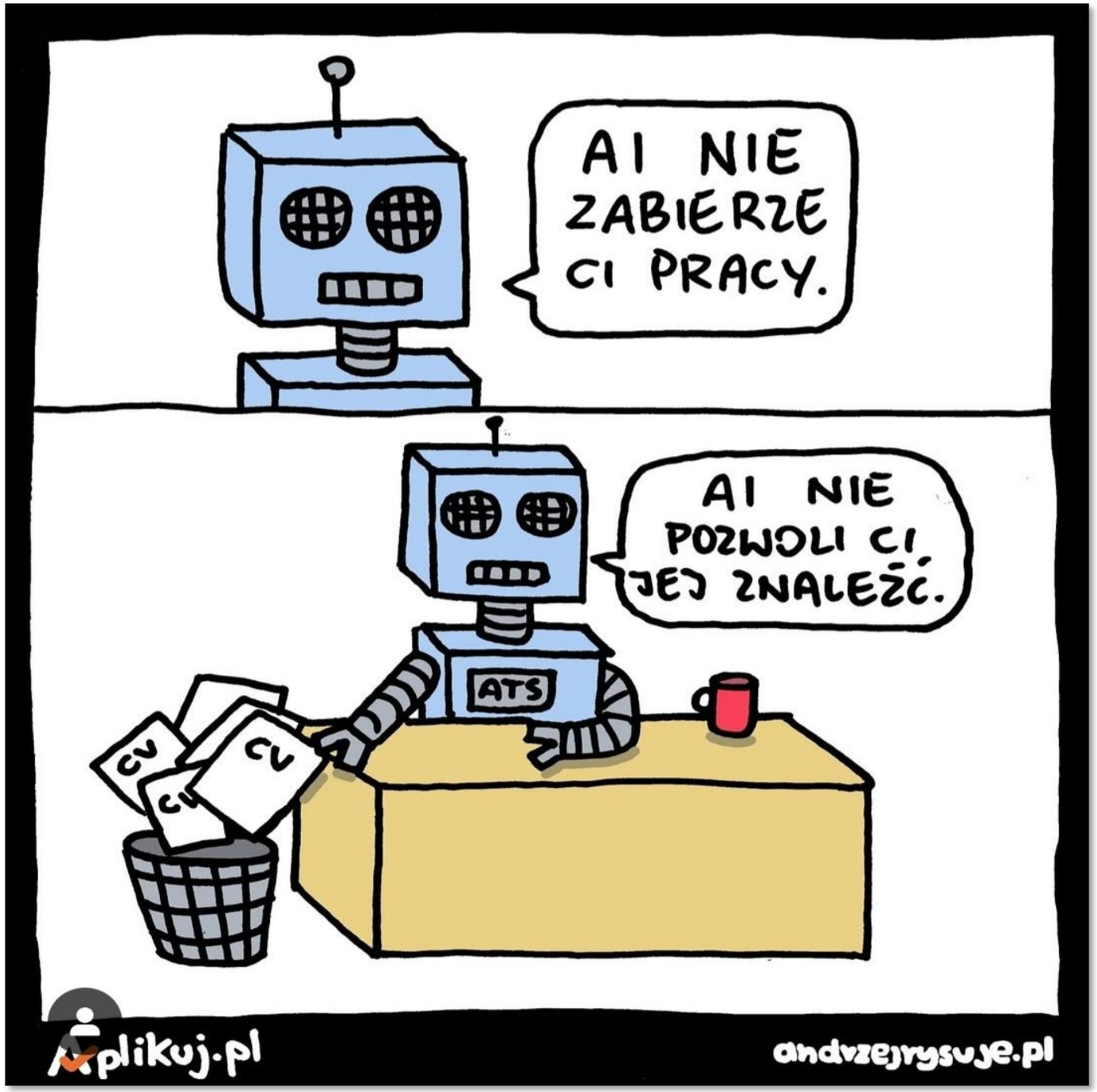
Algorytmy wyszukiwania ścieżek






```
MATCH (n:Miasto {id:7}), (m:Miasto {id:34})
CALL apoc.algo.dijkstra(n, m, 'DROGA', 'km') YIELD path, weight
RETURN path, weight
```

https://neo4j.com/


AI NIE ZABIERZE CI PRACY.
AI NIE POZWOLI CI JEJ ZNALEŹĆ.
ATS
CV
CV
CV
Aplikuj.pl
andrzejrysuje.pl