

Algorytmy Ucznia Maszynowego: Ilustrowany Przewodnik

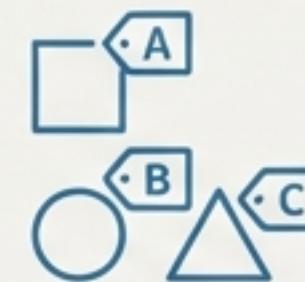
Odkryj swoją skrzynkę z narzędziami Data Science



Przewodnik zaprojektowany do samodzielnej nauki. Przejrzysty, wizualny i bez zbędnego żargonu.

Zacznijmy od mapy: Czym jest Uczenie Maszynowe?

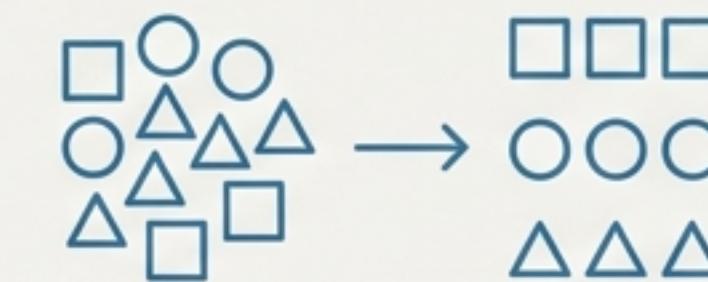
Uczenie maszynowe (ML) to dziedzina sztucznej inteligencji, w której algorytmy uczą się na podstawie danych, aby wykonywać zadania bez jawnych instrukcji. To nauka wzorców, a nie programowanie reguł.



Uczenie Nadzorowane

Algorytm uczy się na podstawie danych z etykietami (tzw. 'prawidłowymi odpowiedziami'). Celem jest przewidywanie tych etykiet dla nowych, nieznanych danych.

Analogia: Pokazujesz dziecku zdjęcia kotów i psów z podpisami, a następnie prosisz, by zidentyfikowało zwierzę na nowym zdjęciu.



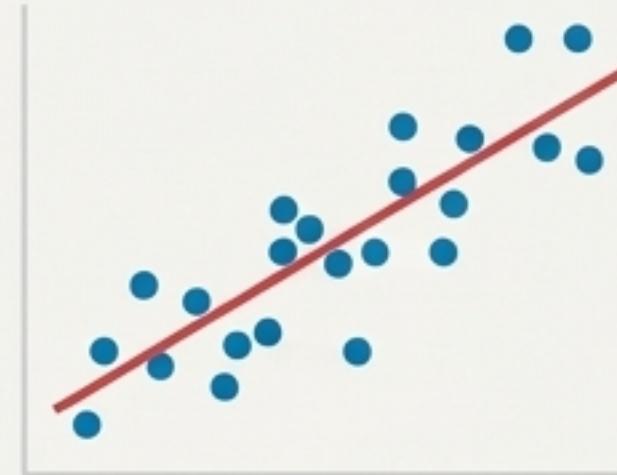
Uczenie Nienadzorowane

Algorytm pracuje na danych bez etykiet. Jego zadaniem jest znalezienie ukrytych struktur, wzorców lub grup (klastrów).

Analogia: Dajesz dziecku stos niepodpisanych zdjęć zwierząt i prosisz, by pogrupowało je według podobieństwa.

W tym przewodniku skupimy się na narzędziach z kategorii uczenia nadzorowanego.

Dwa główne zadania w uczeniu nadzorowanym



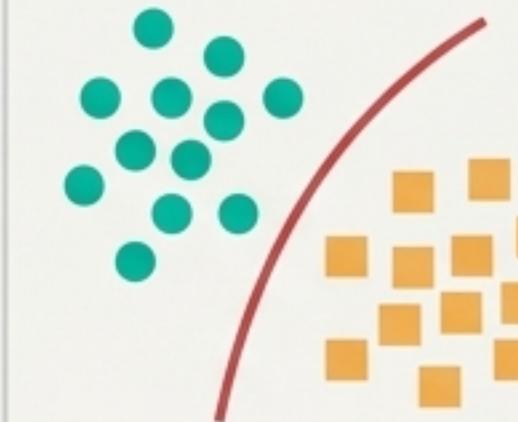
Regresja

Ile? Jaką wartość?

Przewidywanie ciągłej wartości liczbowej.
Szukamy zależności między zmiennymi.

Prognozowanie ceny domu na podstawie jego powierzchni. Cena (np. 500 000 zł, 750 000 zł) jest wartością ciągłą.

Wynik: Liczba (np. 685 340 zł)



Klasyfikacja

Która kategoria? Jaki typ?

Przypisywanie obiektu do jednej z predefiniowanych, dyskretnych klas.

Określenie, czy e-mail to 'spam', czy 'nie spam'. Kategorie są z góry ustalone.

Wynik: Etykieta (np. 'spam')

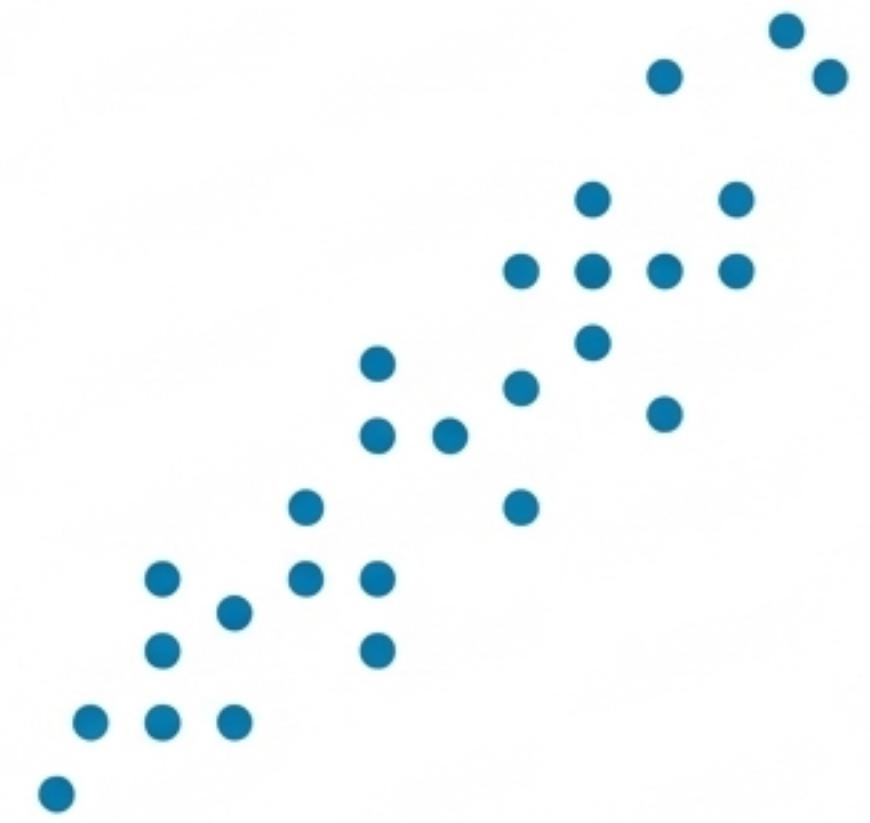


Narzędzie 1: Regresja Liniowa

Problem: Jak przewidzieć cenę mieszkania na podstawie jego metrażu?

Kluczowa Idea: Znajdź prostą linię, która najlepiej opisuje zależność między dwiema zmiennymi, minimalizując średni błąd prognozy.

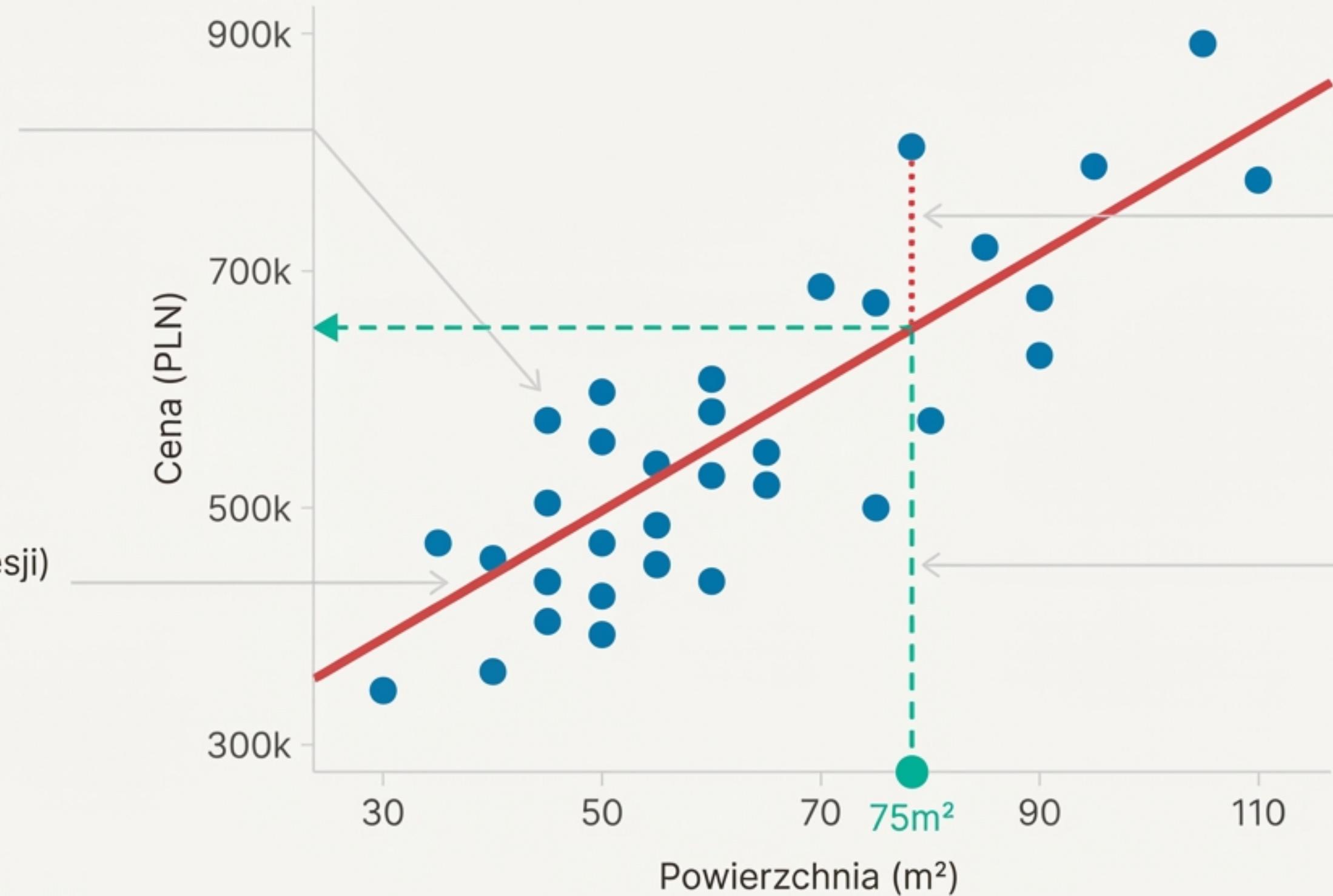
Jak to działa?: Algorytm dopasowuje linię prostą ($y = ax + b$) do danych. Celem jest znalezienie takich parametrów (a, b), aby suma kwadratów odległości wszystkich punktów od tej linii była jak najmniejsza. To minimalizuje błąd.



Regresja Liniowa w praktyce

Dane treningowe
Inter Regular/*Italic*

Model (linia regresji)
Inter Regular/*Italic*



Błąd (rezydua)
Inter Regular/
Algorytm minimalizuje
sumę kwadratów tych
odległości.

Nowa prognoza
Inter Regular/*Italic*
Dla nowej powierzchni
($75m^2$) model prognozuje
cenę ~650k PLN.

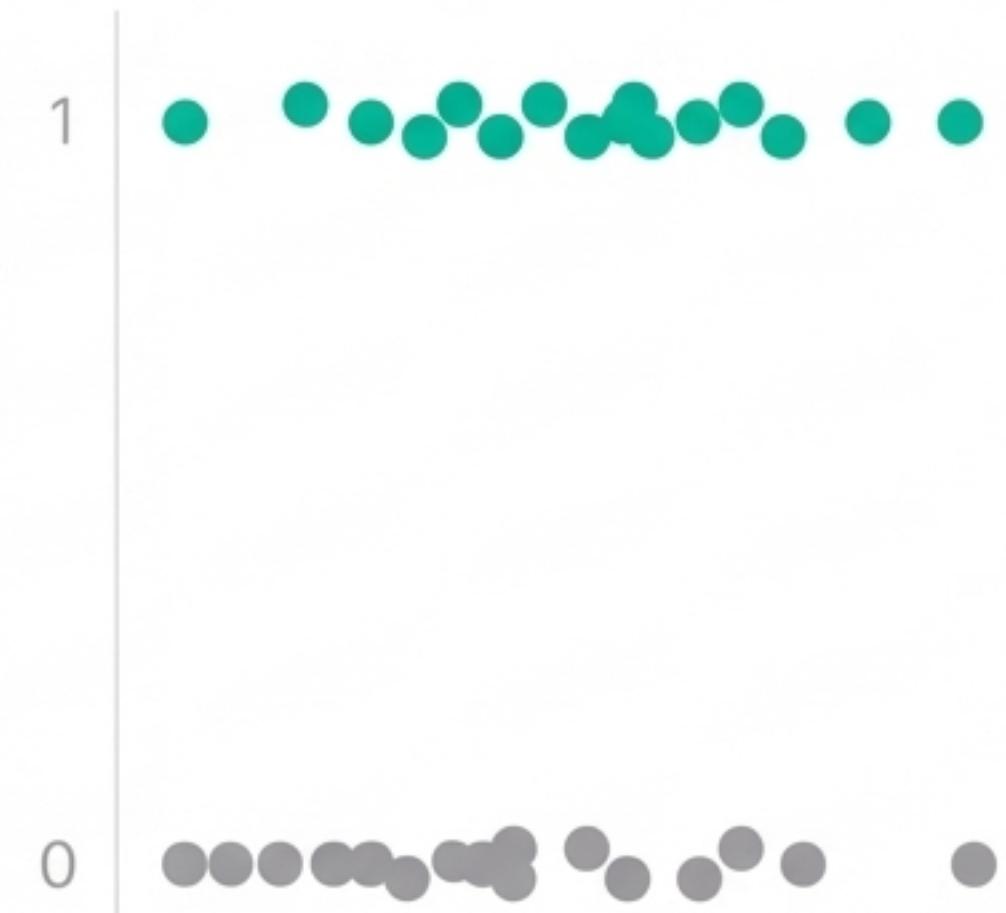


Narzędzie 2: Regresja Logistyczna

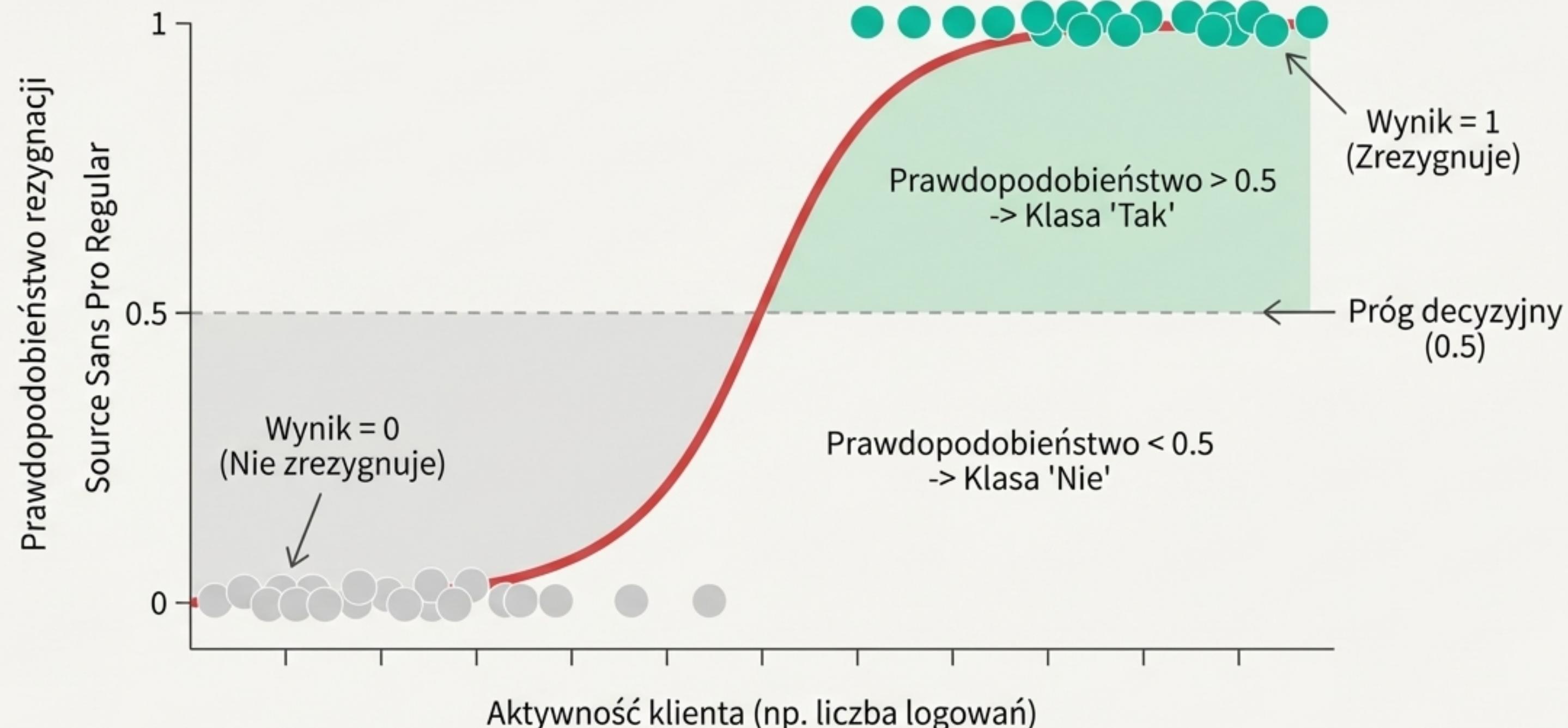
Problem: Czy klient zrezygnuje z usługi na podstawie jego aktywności? (Tak / Nie)

Kluczowa Idea: Zamiast dopasowywać prostą linię, dopasuj funkcję w kształcie litery 'S' (sigmoidę), która 'ścisza' wynik do wartości między 0 a 1, interpretowanej jako prawdopodobieństwo.

Jak to działa?: Model oblicza prawdopodobieństwo przynależności do danej klasy. Jeśli prawdopodobieństwo przekracza ustalony próg (np. 0.5), obiekt jest przypisywany do klasy 'Tak', w przeciwnym razie do klasy 'Nie'.



Regresja Logistyczna w praktyce





Narzędzie 3: K-Najbliższych Sąsiadów (KNN)

Problem (Pytanie)

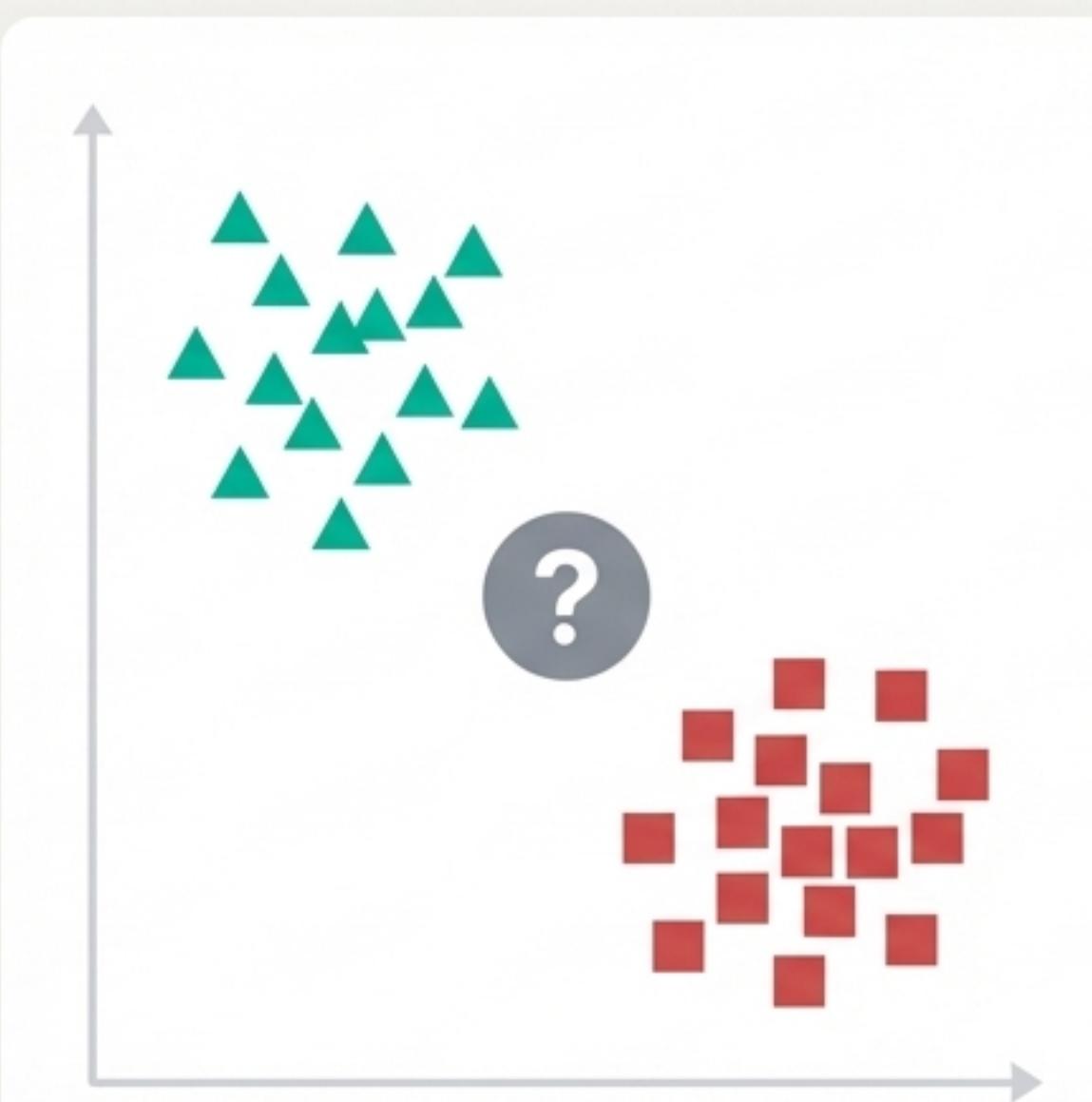
Do jakiej grupy klientów powinniśmy zaklasyfikować nowego użytkownika na podstawie jego wieku i wydatków?

Kluczowa Idea

Jesteś tym, z kim przestajesz. Klasa nowego punktu jest określana przez większość głosów jego 'K' najbliższych sąsiadów.

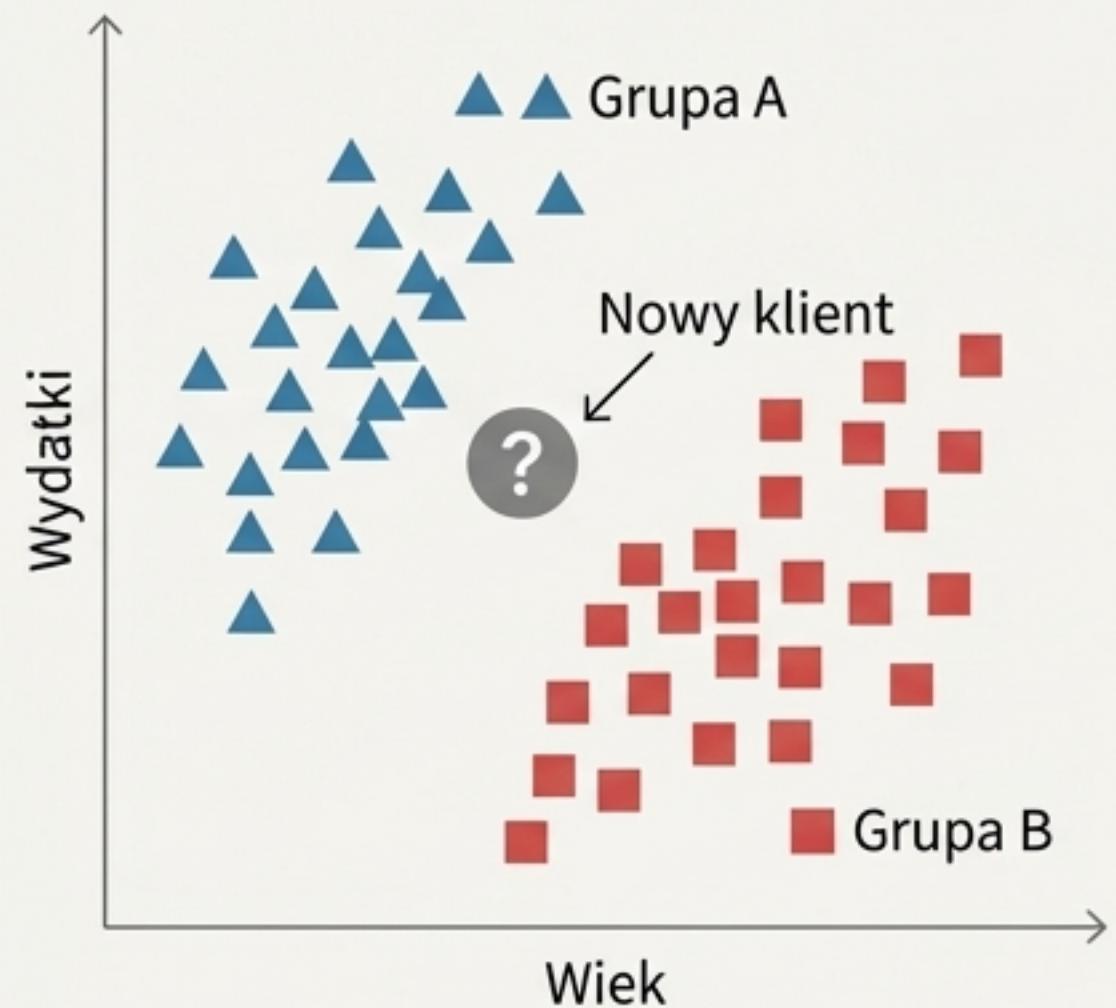
Jak to działa?

To algorytm 'leniwy' (nie-parametryczny) - nie buduje z góry modelu. Aby sklasyfikować nowy punkt, po prostu patrzy na etykiety jego K najbliższych, znanych już punktów i wybiera najczęstszą z nich.



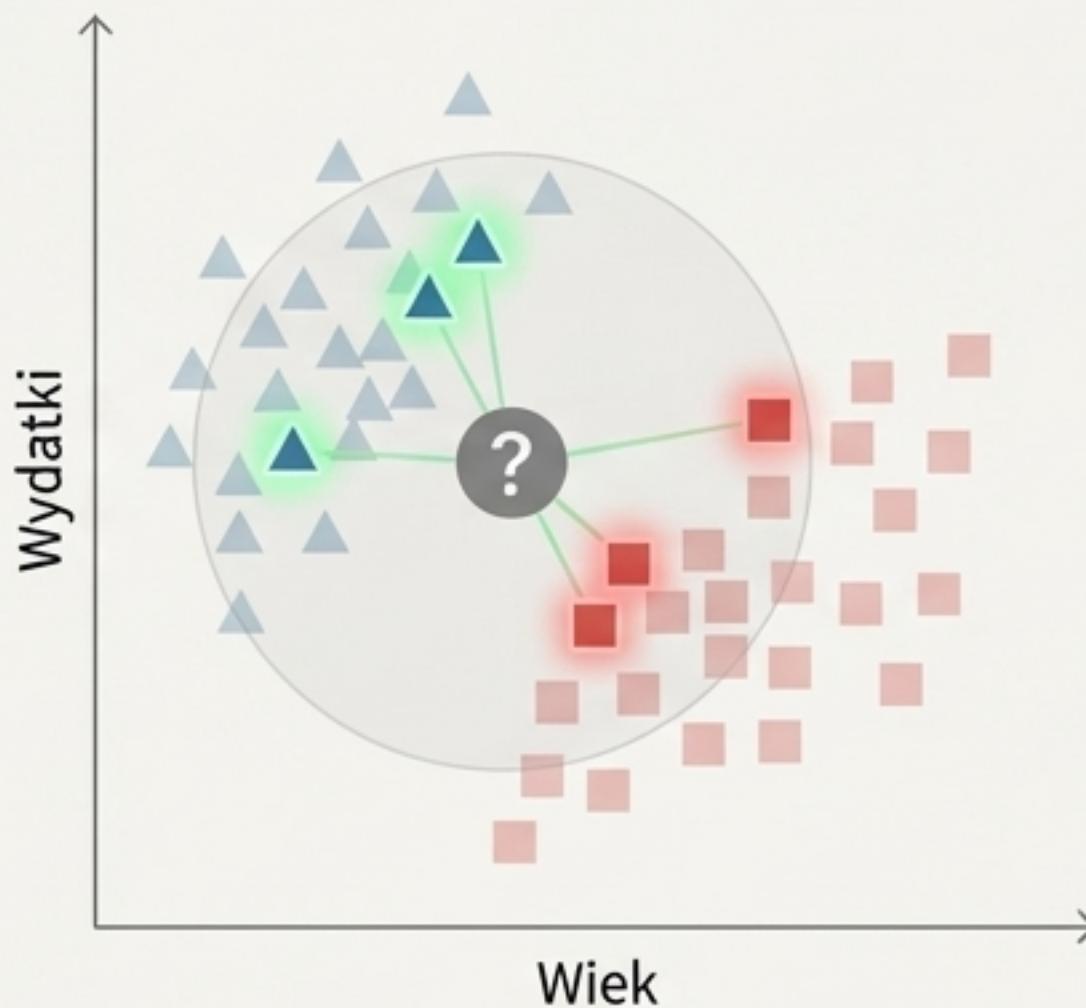
KNN w praktyce (dla K=5)

Krok 1: Dane



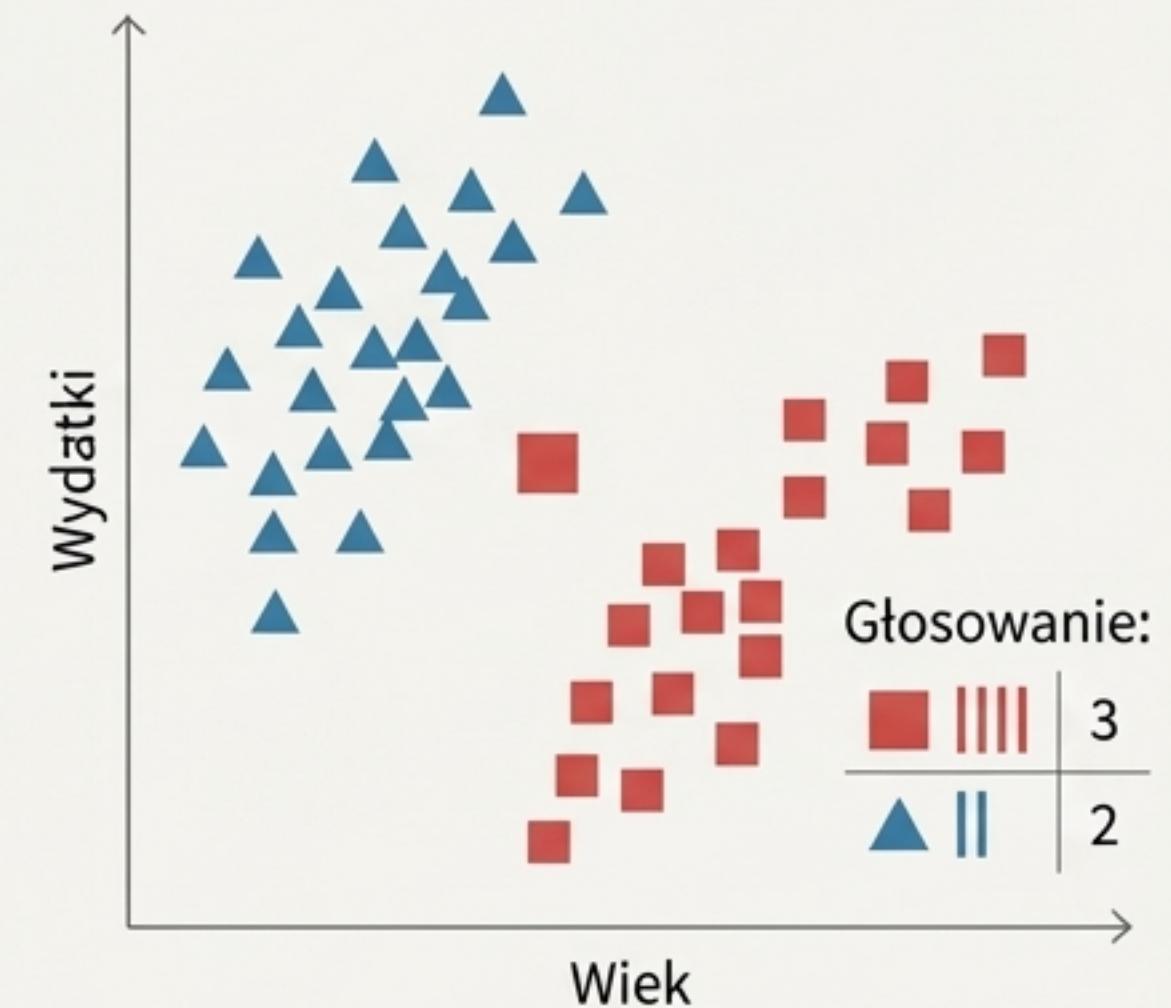
Mamy nowego klienta (?) i chcemy go przypisać do grupy A lub B.

Krok 2: Znajdź K-najbliższych sąsiadów

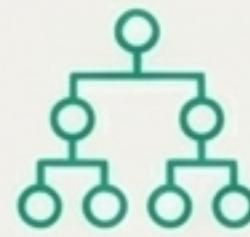


Wybieramy K=5. Algorytm znajduje 5 punktów danych treningowych najbliższej naszego nowego punktu.

Krok 3: Głosowanie większościowe



Większość (3 z 5) sąsiadów należy do grupy B. Nowy punkt zostaje sklasyfikowany jako 'Grupa B'.



Narzędzie 4: Drzewo Decyzyjne

Problem (Pytanie):

Jak zakwalifikować pacjenta do grupy niskiego lub wysokiego ryzyka zawału serca?

Kluczowa Idea:

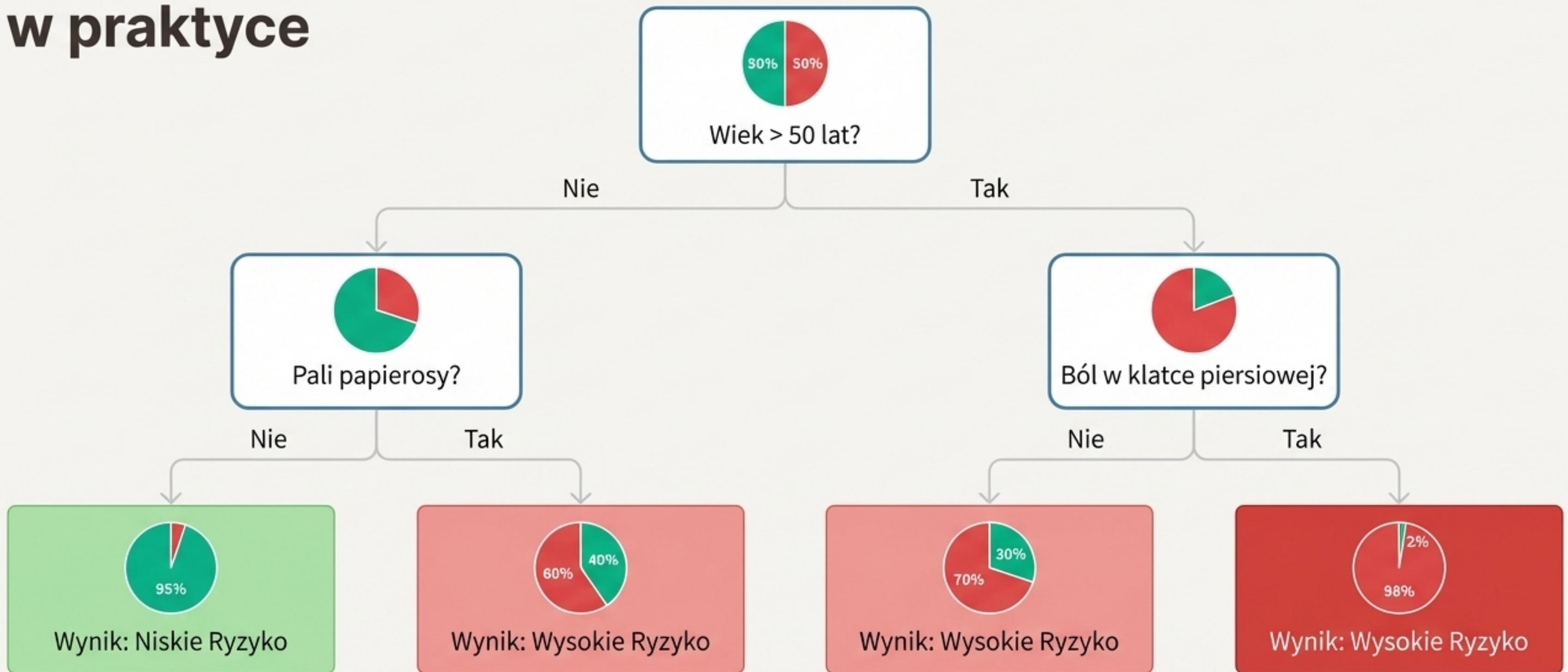
Zbuduj model w formie schematu blokowego, gdzie każdy węzeł to pytanie o cechę, a każda gałąź to odpowiedź, prowadząca do ostatecznej decyzji (liścia).

Jak to działa?:

Algorytm uczy się serii pytań ‘Tak/Nie’, które najlepiej dzielą dane na jak najbardziej jednorodne grupy (‘czyste liście’). Celem jest znalezienie takich podziałów, które maksymalizują czystość grup potomnych.



Drzewo Decyzyjne w praktyce



Celem jest tworzenie "liści", które są jak najbardziej jednorodne (czyste).

A co, jeśli problemem nie jest prognoza, lecz znalezienie najlepszego rozwiązania?

Dotychczasowe narzędzia uczyły się z danych. Istnieje jednak cała klasa problemów optymalizacyjnych, gdzie celem jest znalezienie najlepszej możliwej odpowiedzi spośród miliardów kombinacji.

Przykład: Jaką najkrótszą trasę powinien pokonać kurier, by odwiedzić 20 miast i wrócić do punktu startu?

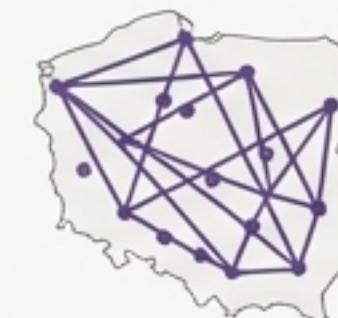


Narzędzia inspirowane naturą

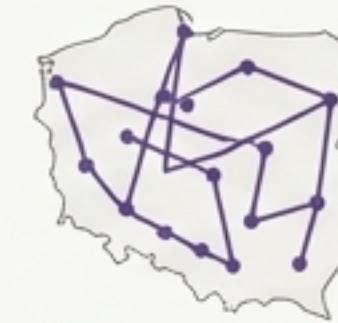
Wprowadzenie do algorytmów ewolucyjnych.



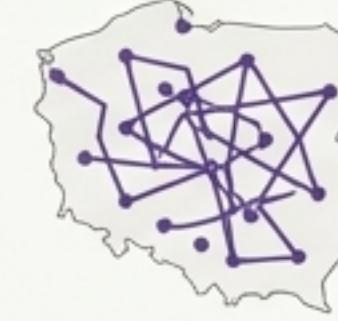
Inny rodzaj narzędzia: Algorytmy Genetyczne



Osobnik 1 /
Fitness: 2450 km



Osobnik 2 /
Fitness: 2610 km



Osobnik 3 /
Fitness: 2380 km

Problem: Jak znaleźć optymalną (najkrótszą) trasę dla komiwojażera? (Problem Komiwojażera - TSP)

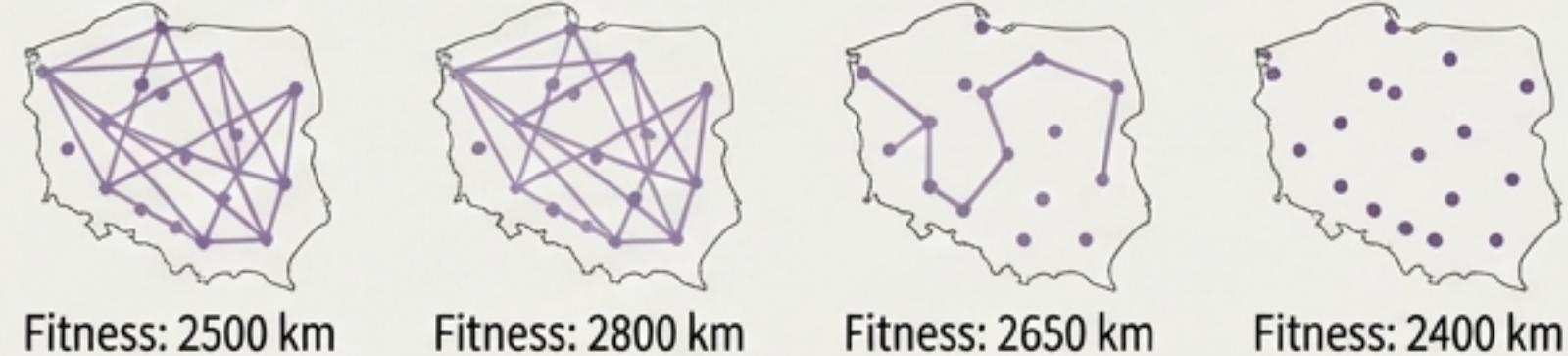
Kluczowa Idea: Zainspirowany ewolucją biologiczną, algorytm generuje populację możliwych rozwiązań (tras) i 'ewoluje' ją przez kolejne pokolenia, pozwalając przetrwać i rozmnażać się tylko najlepszym.

Podstawowe pojęcia:

- **Osobnik (Indywidualny):** Jedno konkretne rozwiązanie (jedna możliwa trasa).
- **Populacja:** Zbiór wielu osobników (wiele różnych tras).
- **Funkcja Przystosowania (Fitness):** Miara jakości rozwiązania (np. całkowita długość trasy - im krótsza, tym lepsza).
- **Ewolucja:** Proces selekcji, krzyżowania i mutacji.

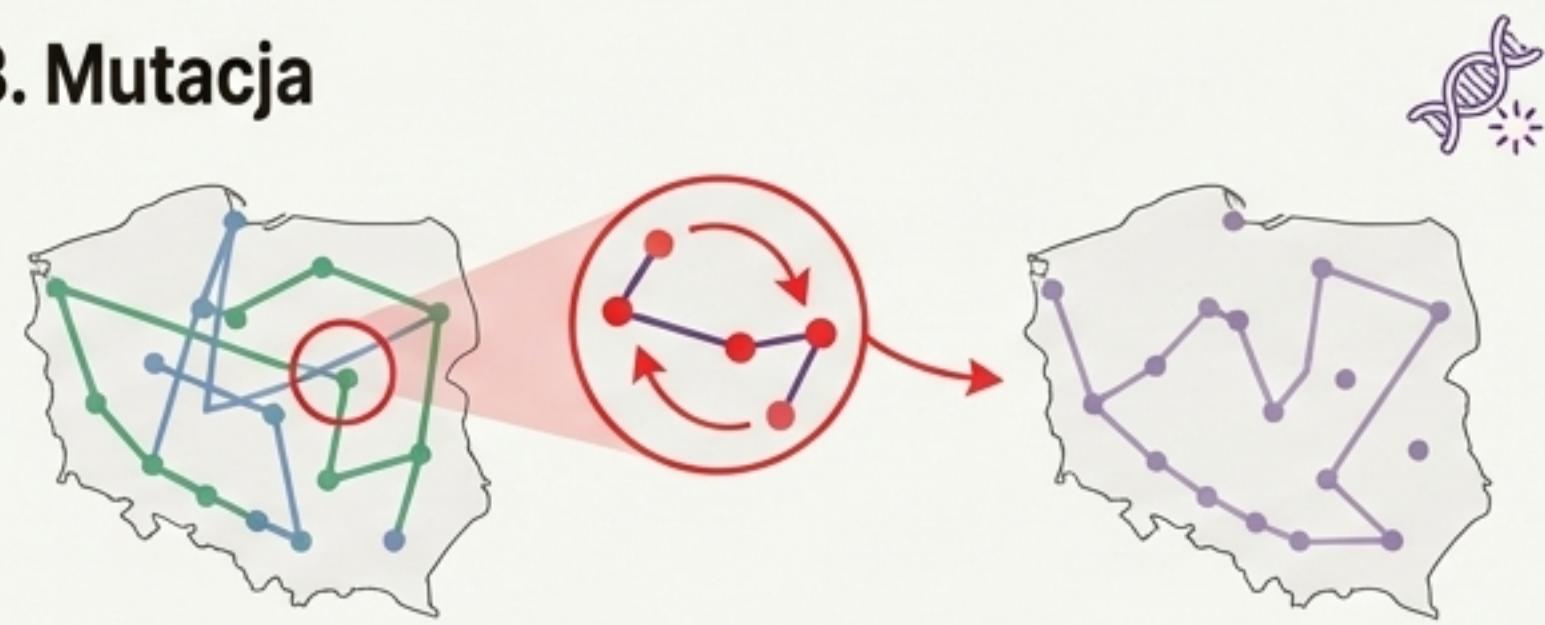
Ewolucja rozwiązań: Algorytm Genetyczny w akcji

1. Populacja początkowa



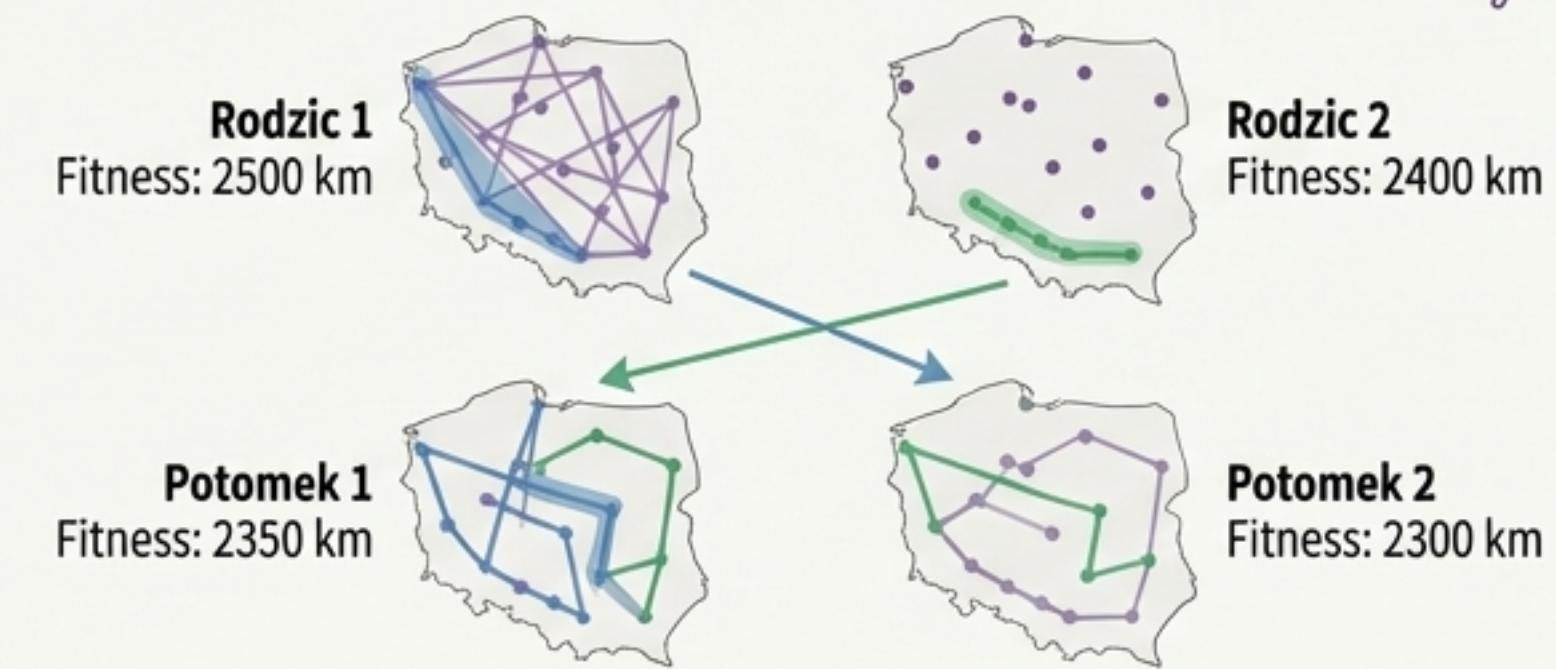
Generujemy losowy zbiór tras.

3. Mutacja



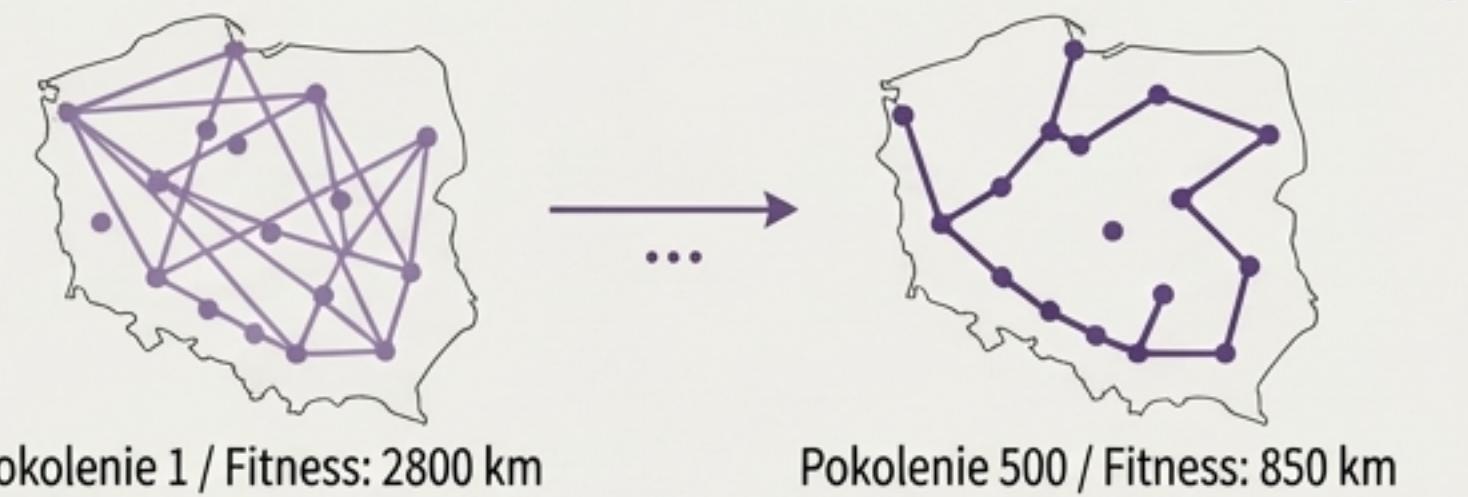
Wprowadzamy losowe, małe zmiany (mutacje), aby zapobiec utknięciu w lokalnym optimum i odkryć nowe możliwości.

2. Selekcja i Krzyżowanie (Crossover)



Najlepsze trasy są wybierane i "krzyżowane", wymieniając fragmenty, by stworzyć nowe, potencjalnie lepsze rozwiązania.

4. Nowe pokolenie i rozwiązanie końcowe



Proces powtarza się przez wiele pokoleń, aż rozwiązanie przestaje się znacząco poprawiać.



Twoja skrzynka z narzędziami jest gotowa



Regresja Liniowa

Do prognozowania wartości liczbowych.



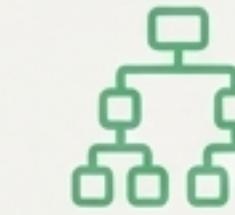
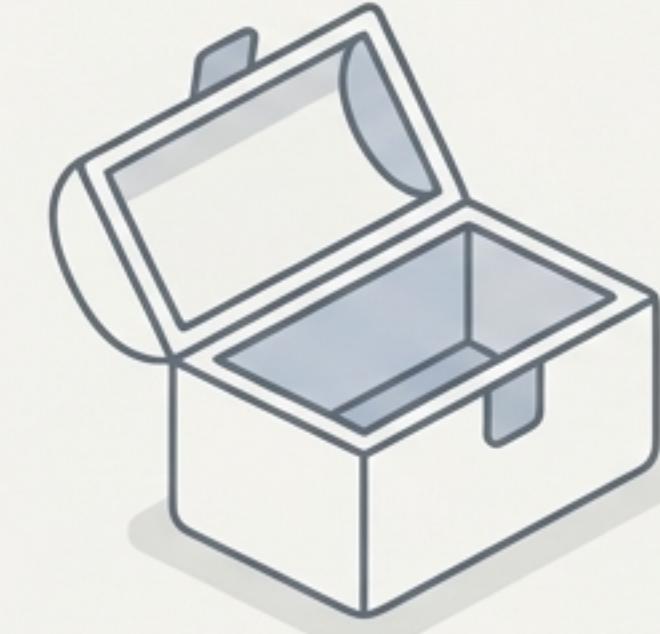
Regresja Logistyczna

Do klasyfikacji binarnej (Tak/Nie).



K-Najbliższych Sąsiadów

Do klasyfikacji opartej na "sąsiedztwie".



Drzewo Decyzyjne

Do tworzenia zrozumiałych, opartych na regulach modeli decyzyjnych.



Algorytmy Genetyczne

Do poszukiwania optymalnych rozwiązań złożonych problemów.

Zrozumienie, ***jak*** działa algorytm, jest pierwszym krokiem do skutecznego rozwiązania problemu. Nie traktuj ich jak czarnych skrynek – traktuj je jak specjalistyczne narzędzia w swoim arsenale.