

Regresja Logistyczna

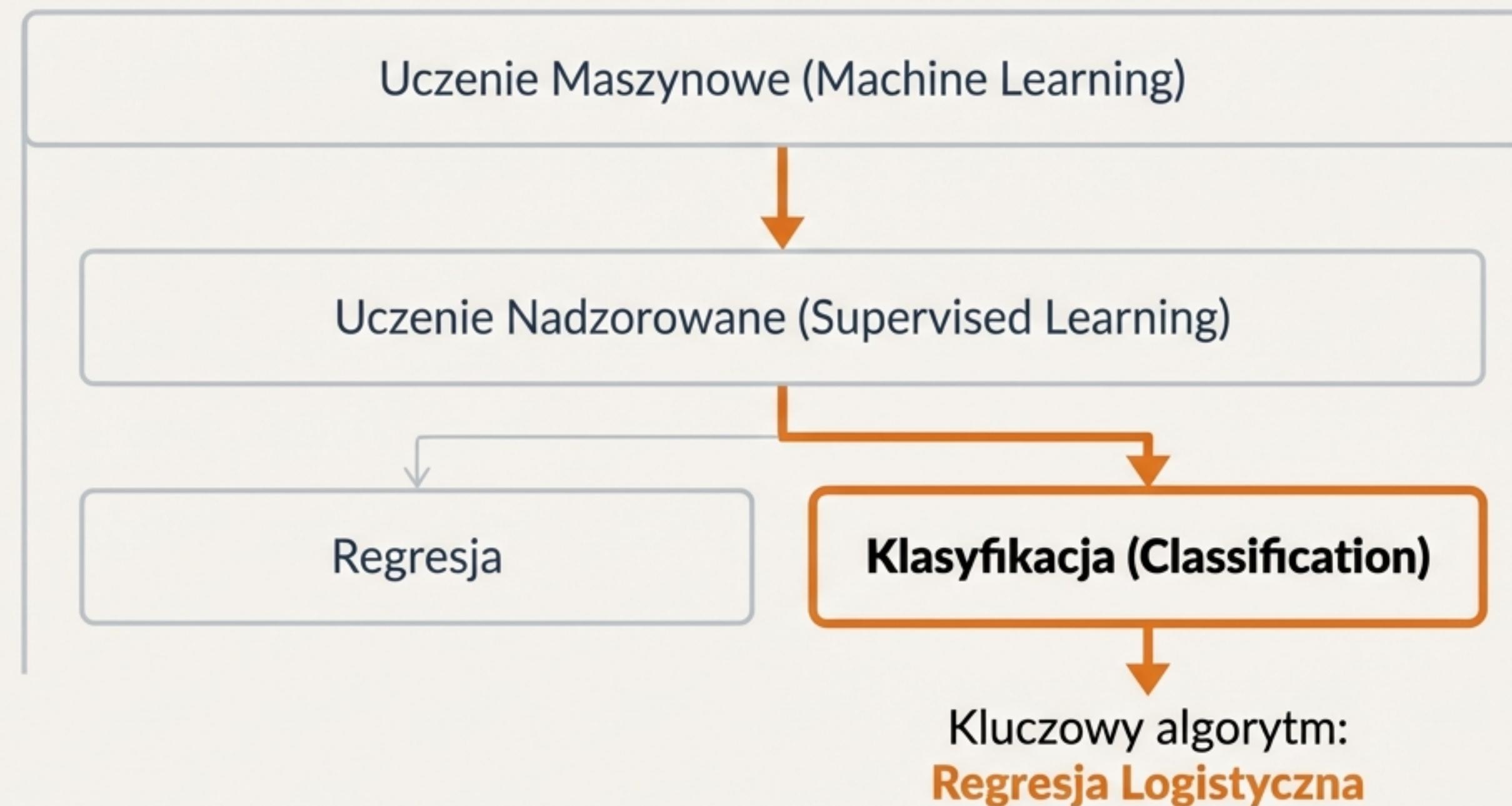
Od Linii Prostej do Prawdopodobieństwa: Dogłębne
Wyjaśnienie Kluczowego Algorytmu Klasyfikacji



Prezentacja typu 'Explainer / Deep Dive'

Gdzie w Świecie Machine Learningu Znajduje Się Regresja Logistyczna?

Regresja logistyczna to fundamentalny algorytm **uczenia nadzorowanego**, wykorzystywany do zadań **klasyfikacji**. Zanim zagłębimy się w jej działanie, umiejscówmy ją na mapie.

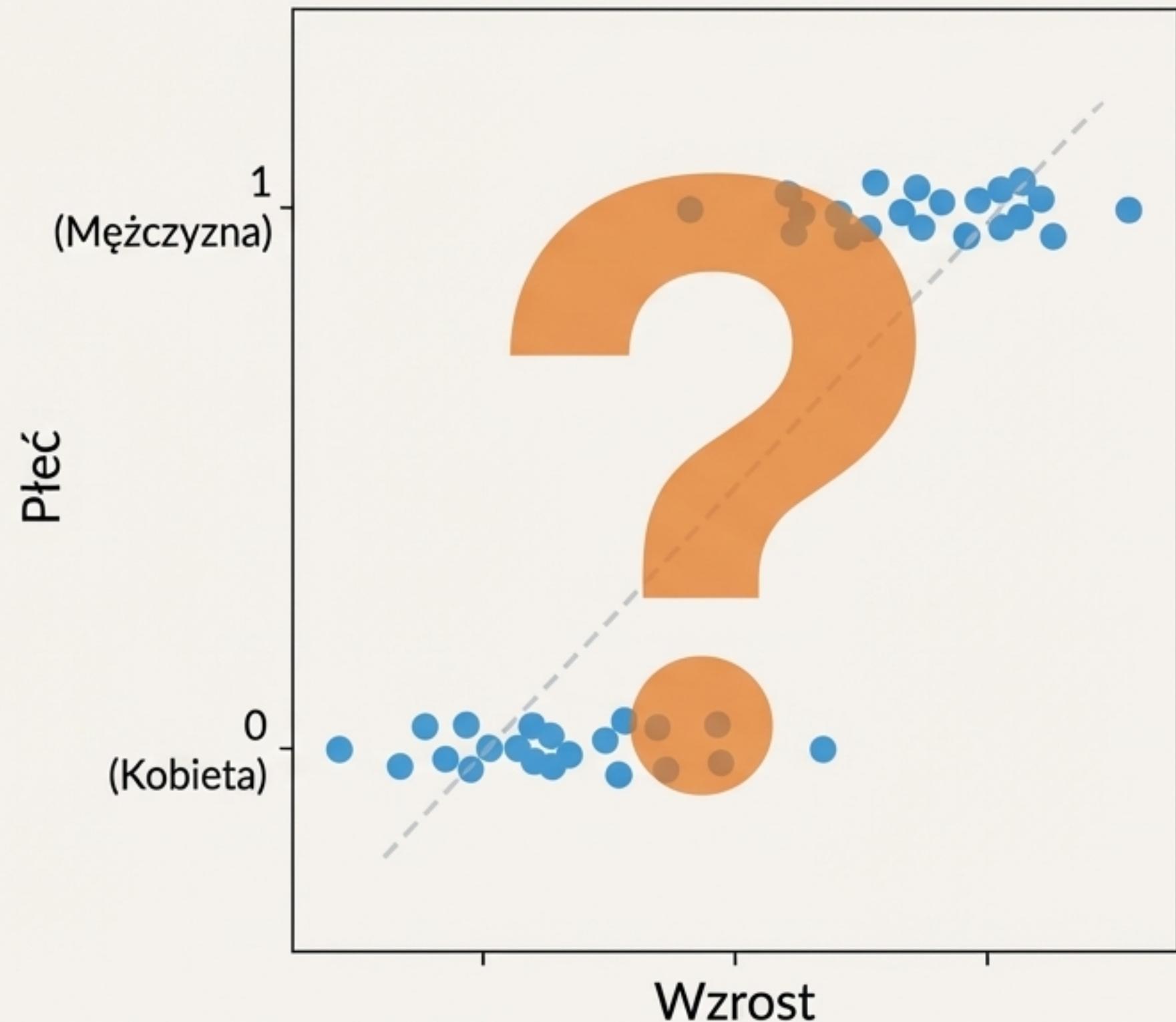


Zdefiniujmy Problem: Klasyfikacja Binarna

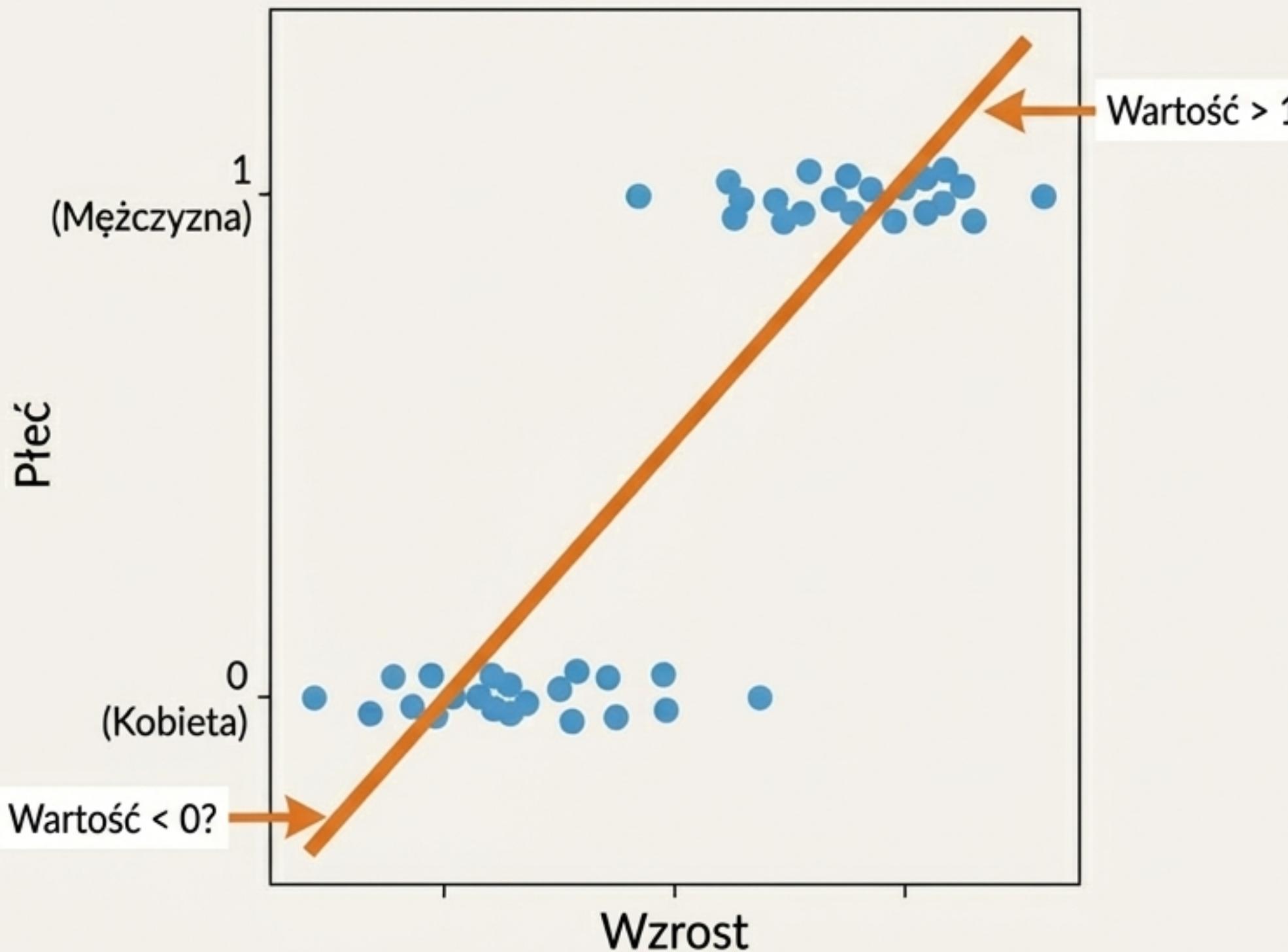
Wyobraźmy sobie typowe zadanie: na podstawie cech (np. wzrostu i wagi) chcemy przewidzieć, czy dana osoba należy do jednej z dwóch kategorii – na przykład, czy jest mężczyzną czy kobietą.

- **Cel:** Przypisać etykietę 'Klasa 0' lub 'Klasa 1'.

Pytanie: Znamy już regresję liniową, 'matkę wszystkich algorytmów ML', która świetnie modeluje zależności liczbowe. Czy moglibyśmy ją zastosować tutaj?



Dlaczego Regresja Liniowa Nie Jest Dobrym Rozwiązaniem?



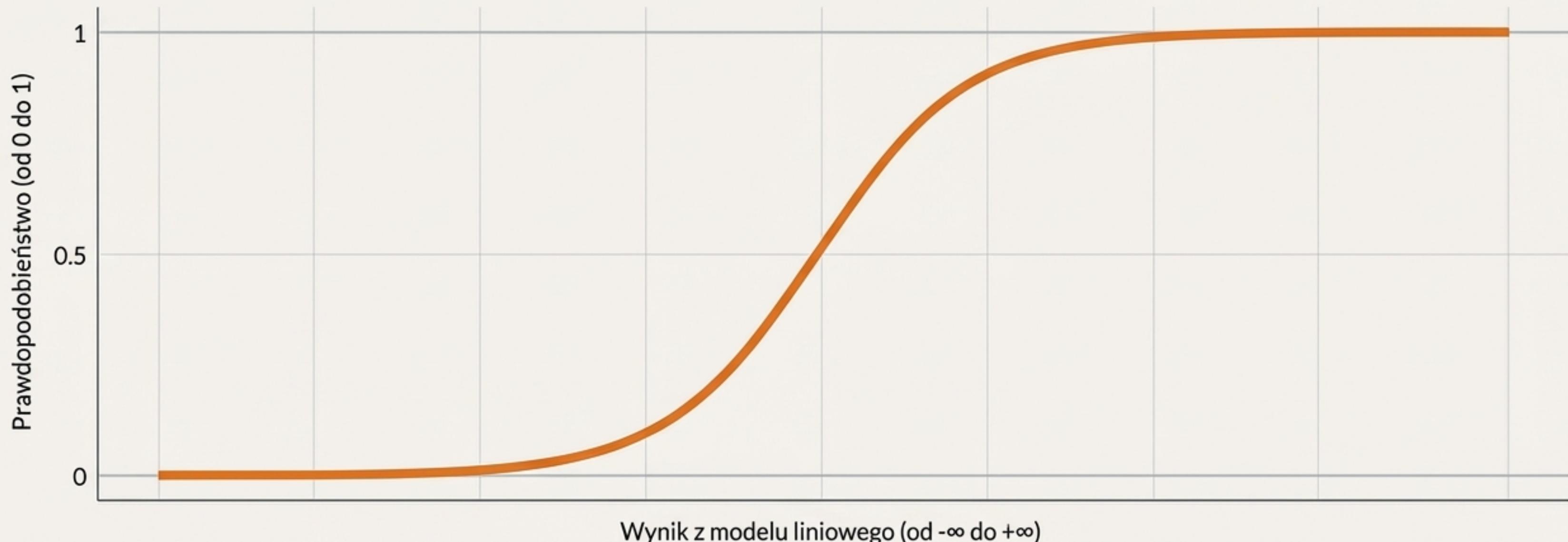
Próba dopasowania prostej linii do problemu klasyfikacyjnego prowadzi do fundamentalnych problemów:

- **1. Wartości poza zakresem:** Linia regresji wykracza daleko poza sensowny przedział $[0, 1]$. Co oznaczałaby 'płeć' o wartości 1.5 lub -0.2?
- **2. Brak interpretacji probabilistycznej:** Wynik nie jest prawdopodobieństwem. To po prostu liczba na nieskończonej linii, co utrudnia podjęcie decyzji.

Rozwiążanie: Elegancka Krzywa Zamiast Prostej Linii

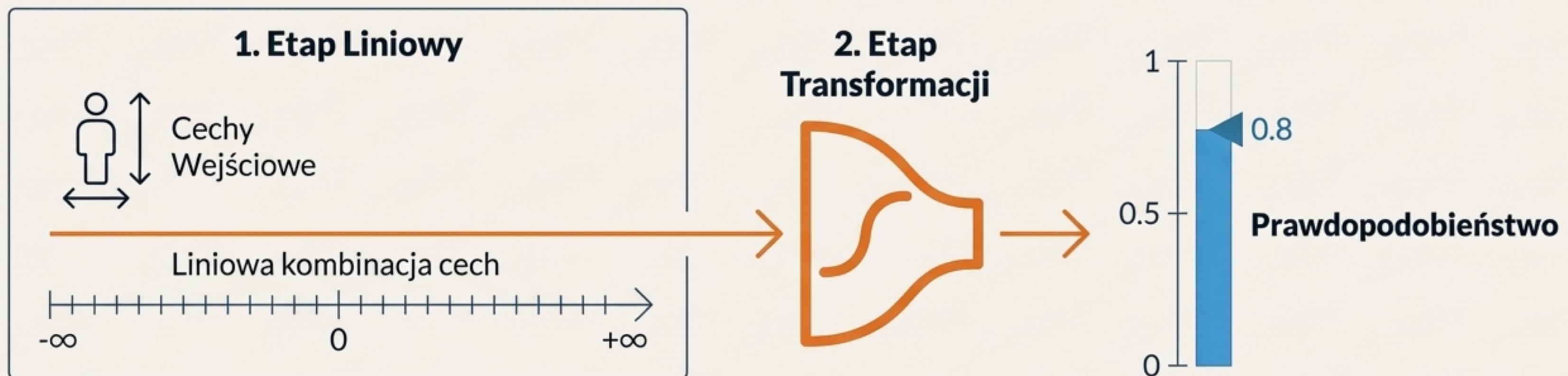
Regresja logistyczna porzuca linię prostą na rzecz **funkcji sigmoidalnej** (lub logistycznej). Ta krzywa w kształcie litery 'S' jest matematycznie zaprojektowana, aby rozwiązać problemy, z którymi nie radzi sobie regresja liniowa.

Jej kluczową cechą jest to, że bez względu na wartość wejściową, jej wynik **zawsze** mieści się w przedziale od 0 do 1.



Genialny Mechanizm: "Zgniatanie" Nieskończoności do Prawdopodobieństwa

Sedno regresji logistycznej to dwuetapowy proces:



Wewnątrz, algorytm wciąż oblicza wynik w oparciu o liniową kombinację cech... Ten wynik może przyjąć dowolną wartość.

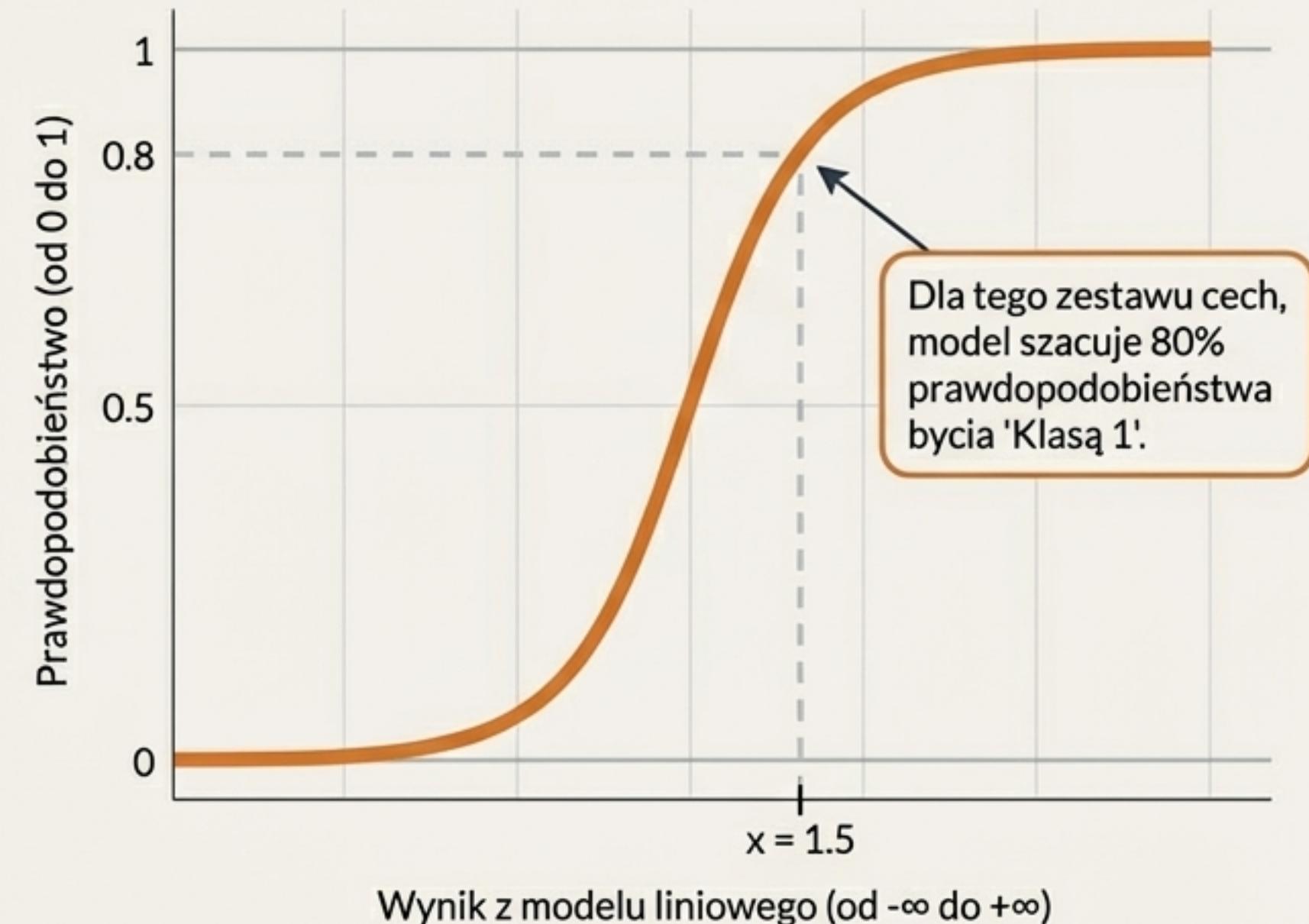
Ten "surowy" wynik jest następnie przepuszczany przez funkcję sigmoidalną, która "zgniata" całą oś liczbową do zgrabnego przedziału [0, 1].

Jak Interpretować Wynik? To Jest Prawdopodobieństwo.

- Wartość wyjściowa z funkcji sigmoidalnej to nic innego jak **prawdopodobieństwo**, że dany punkt danych należy do 'klasy 1'.
- Wynik **0.8** oznacza 80% prawdopodobieństwa przynależności do klasy 1.
- Wynik **0.1** oznacza 10% prawdopodobieństwa przynależności do klasy 1 (i 90% do klasy 0).

Przykład z życia (konsepcja z materiału źródłowego):

> "Prawdopodobieństwo, że dorosła osoba o wzroście 180 cm jest mężczyzną, wynosiłoby 80%."



Od Prawdopodobieństwa do Decyzji: Granica Decyzyjna

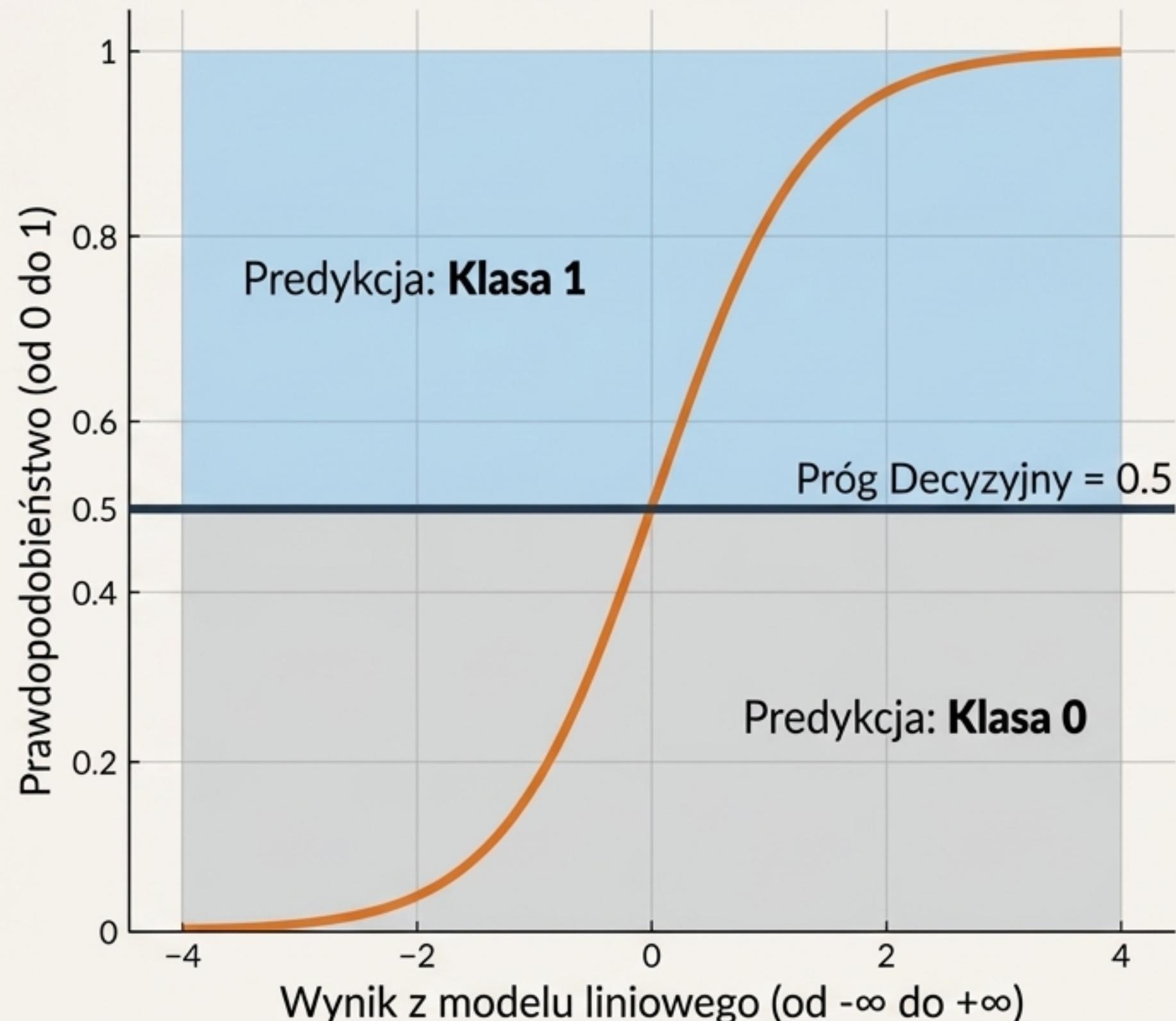
Aby dokonać ostatecznej klasyfikacji, potrzebujemy progu decyzyjnego.

Standardowo przyjmuje się wartość 0.5.

Jeśli $P(\text{Klasa 1}) \geq 0.5$, przypisujemy obiekt do Klasa 1.

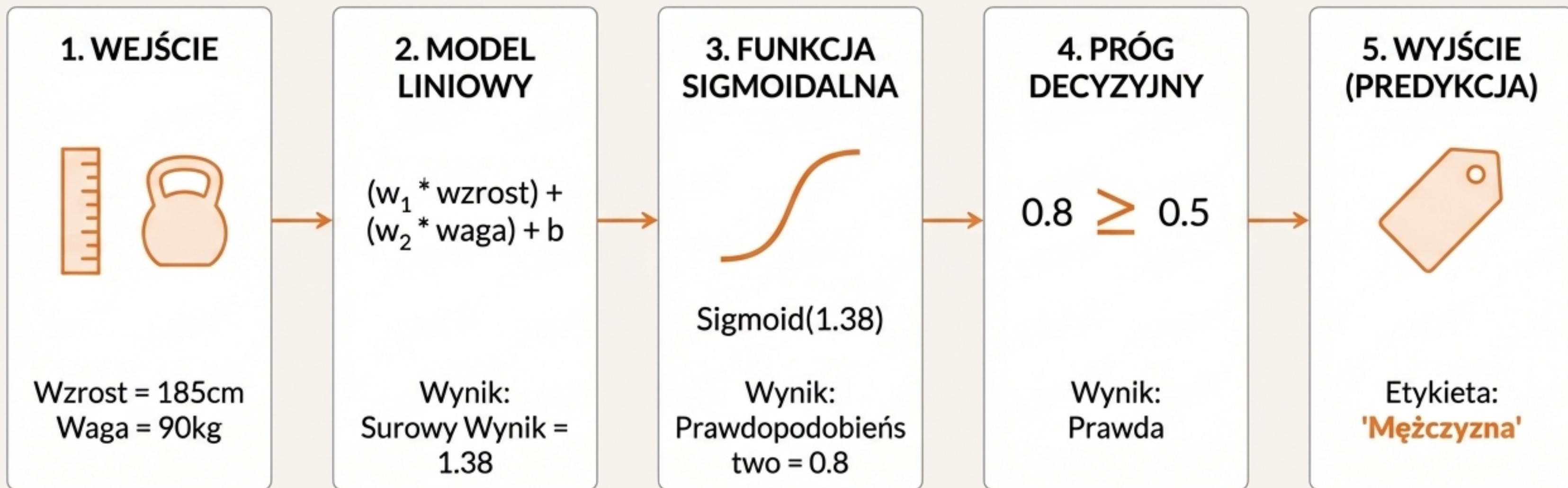
Jeśli $P(\text{Klasa 1}) < 0.5$, przypisujemy obiekt do Klasa 0.

Próg ten można dostosować w zależności od problemu biznesowego (np. w diagnostyce medycznej możemy chcieć być bardziej ostrożni i obniżyć próg, aby nie przegapić chorego pacjenta).



Cały Proces w 5 Krokach: Przykład

Zobaczmy, jak to działa na przykładzie przewidywania płci (**Klasa 1 = Mężczyzna**) na podstawie wzrostu i wagi.



Ciekawostka: Skąd Nazwa "Logistyczna"?

Nazwa pochodzi od funkcji 'logit', która jest odwrotnością funkcji sigmoidalnej. Funkcja logit przekształca prawdopodobieństwo (z przedziału 0-1) z powrotem na pełną oś liczbową. Regresja logistyczna modeluje **logarytm szansy** (log-odds) jako funkcję liniową cech wejściowych. To pokazuje, jak głęboko jest ona zakorzeniona w mechanizmie regresji liniowej.



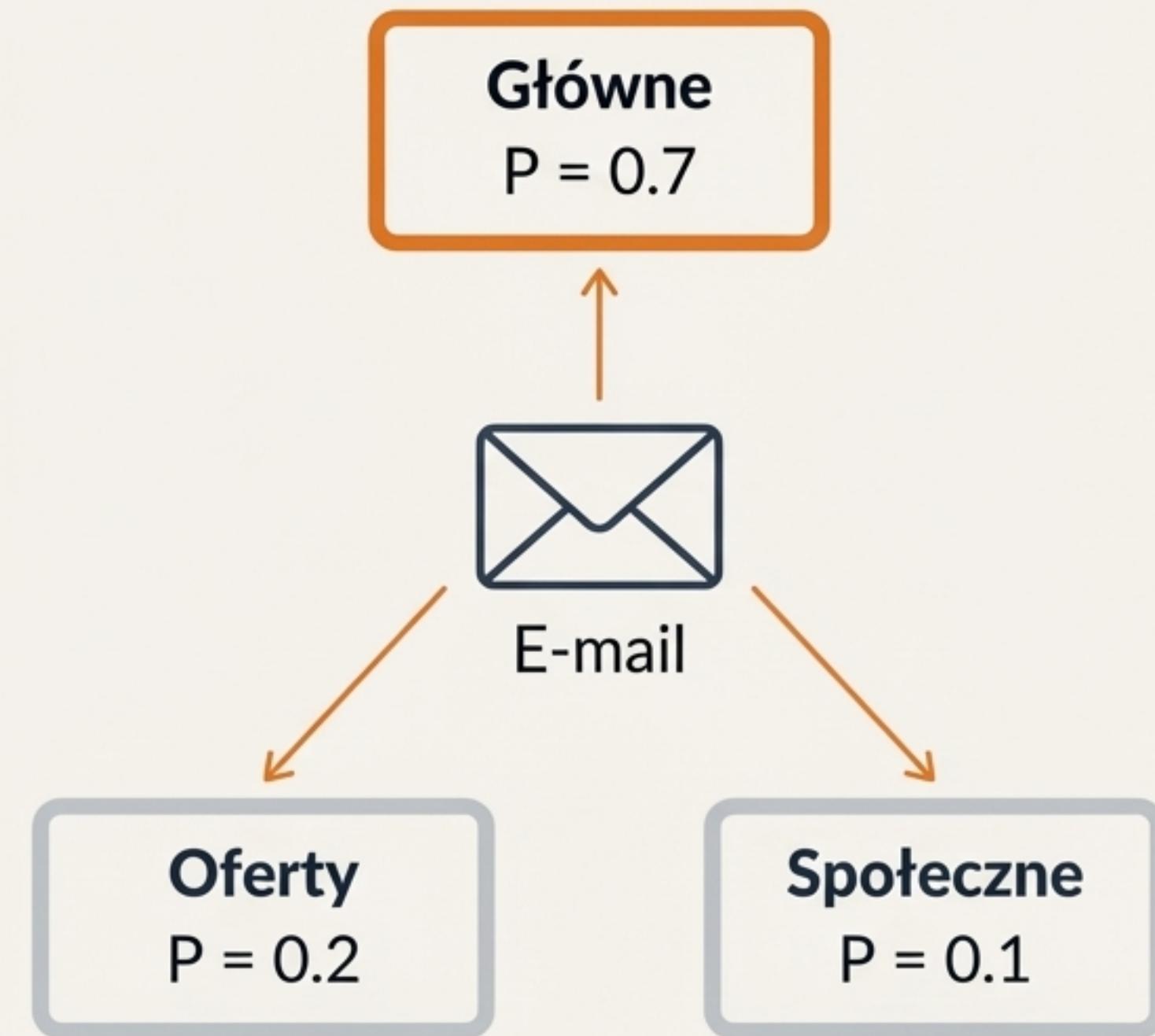
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \dots$$

Co, Jeśli Mamy Więcej Niż Dwie Klasy?

Regresja logistyczna nie jest ograniczona do problemów binarnych. Jej rozszerzenie, znane jako **Regresja Logistyczna Wielomianowa** (**Multinomial Logistic Regression**), potrafi obsługiwać wiele klas.

Zamiast jednej funkcji sigmoidalnej, model tworzy funkcję dla każdej klasy i normalizuje wyniki (za pomocą funkcji softmax), aby uzyskać rozkład prawdopodobieństwa.

Przykład: Automatyczna kategoryzacja e-maili w Gmailu na 'Główne', 'Społeczne', 'Oferty' itd.



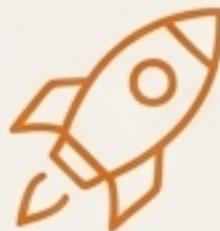
Mocne Strony i Idealne Zastosowania

Kiedy warto użyć Regresji Logistycznej?

Kluczowe Zalety



Interpretowalność: Współczynniki modelu mówią o wpływie każdej cechy na wynik.



Wydajność: Jest szybka w trenowaniu i nie wymaga dużych zasobów obliczeniowych.



Zwraca Prawdopodobieństwa: Zamiast samej etykiety, dostajemy ocenę pewności modelu.



Doskonały Model Bazowy: To świetny punkt startowy do porównywania z bardziej złożonymi algorytmami.

Idealne Scenariusze

- Problemy klasyfikacji binarnej (np. ocena ryzyka kredytowego, detekcja spamu, churn klienta).
- Gdy potrzebna jest szybka i interpretowalna pierwsza wersja modelu.
- Jako komponent w bardziej złożonych systemach.

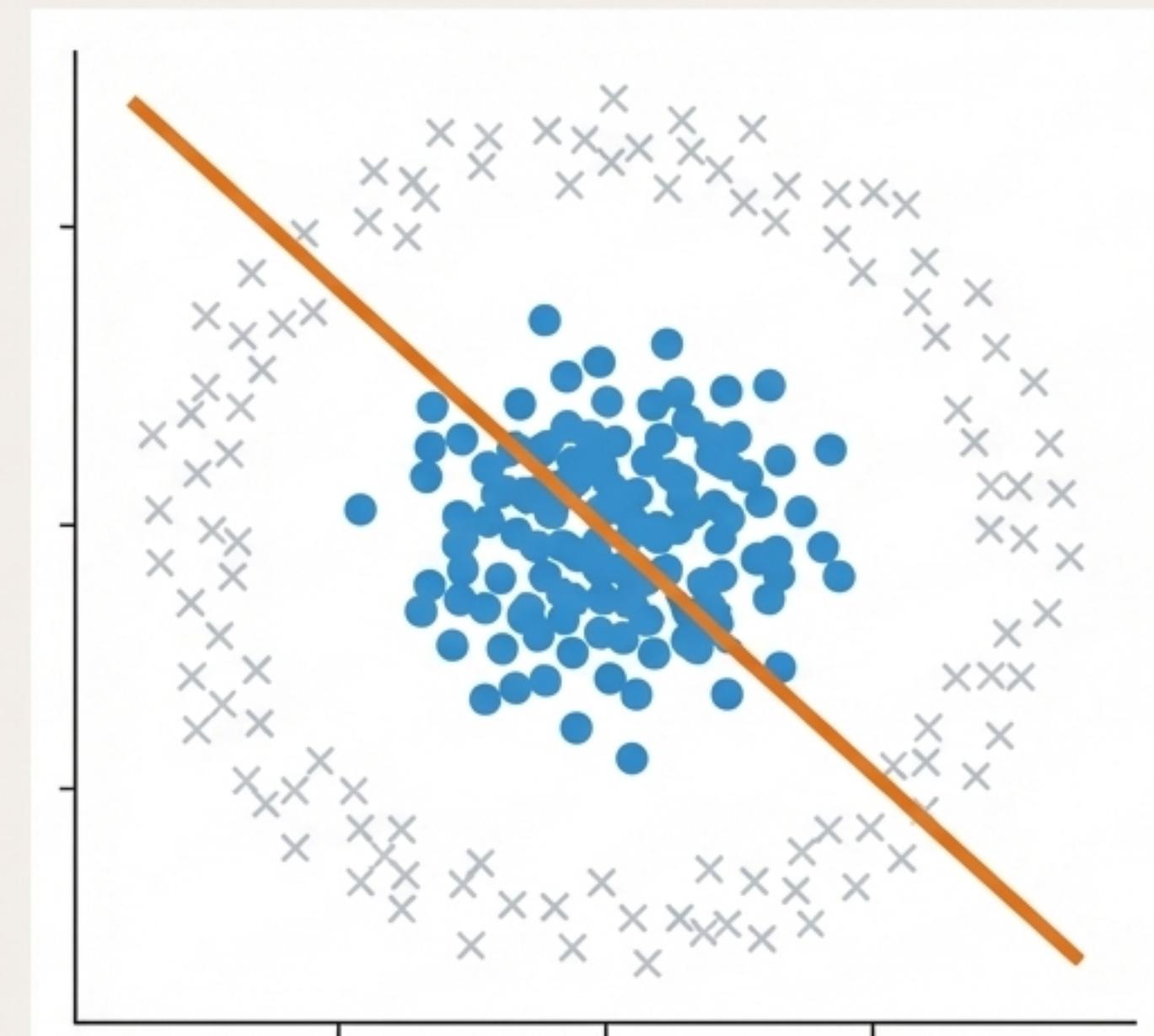
Ograniczenia i Rzeczy, o Których Należy Pamiętać

Mimo swojej użyteczności, regresja logistyczna ma swoje ograniczenia.

Założenie o Liniowości: Model zakłada, że istnieje liniowa zależność między cechami a logarytmem szansy. Jeśli zależność jest bardziej skomplikowana, model może sobie nie poradzić.

Wrażliwość na Wartości Zależne: Słabo radzi sobie z silnie skorelowanymi cechami.

Wydajność na Złożonych Danych: Często jest prześcigana przez bardziej zaawansowane modele (jak SVM z jądrami czy sieci neuronowe) na problemach z nieliniowymi granicami decyzyjnymi.



Fundament, na Którym Stoją Giganci

Zrozumienie regresji logistycznej to więcej niż nauka jednego algorytmu. To zrozumienie fundamentalnego konceptu, który jest podstawą znacznie bardziej złożonych architektur.

Pojedynczy neuron w **sieci neuronowej** często działa w bardzo podobny sposób: przyjmuje ważoną sumę wejść i przepuszcza ją przez funkcję aktywacji (często podobną do sigmoidalnej), aby wygenerować sygnał wyjściowy. Jak ujął to autor materiału źródłowego: “aby zrozumieć sieci neuronowe, spójrzmy ponownie na regresję logistyczną.”

