



**GHOST**

Group of Horribly Optimistic Statisticians



# Intro to ML

## #4 Klasyfikacja





**GHOST**

Group of Horribly Optimistic Statisticians



# Agenda

1. Regresja - przypomnienie
2. Co to jest klasyfikacja?
3. Metody oceny klasyfikacji
4. Proste klasyfikatory
5. Wyzwania



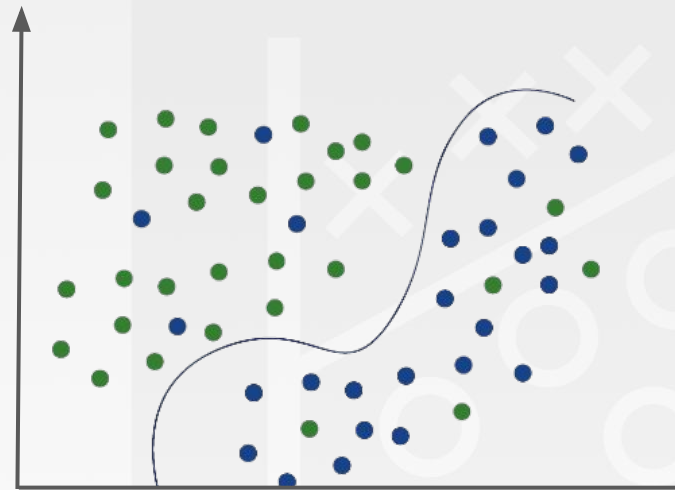
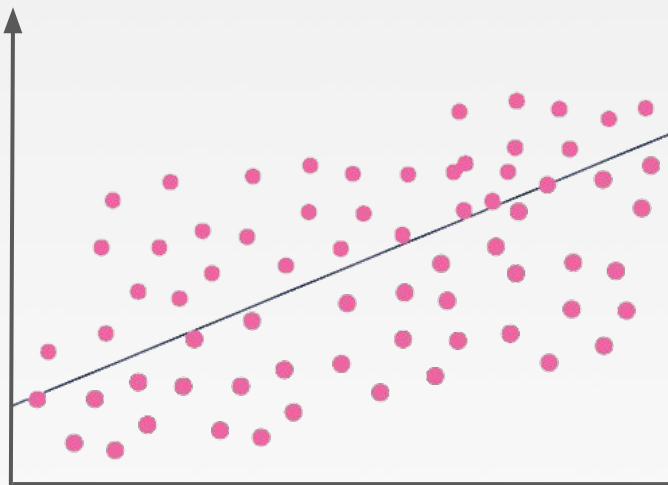


**GHOST**

Group of Horribly Optimistic Statisticians



# Regresja vs klasyfikacja



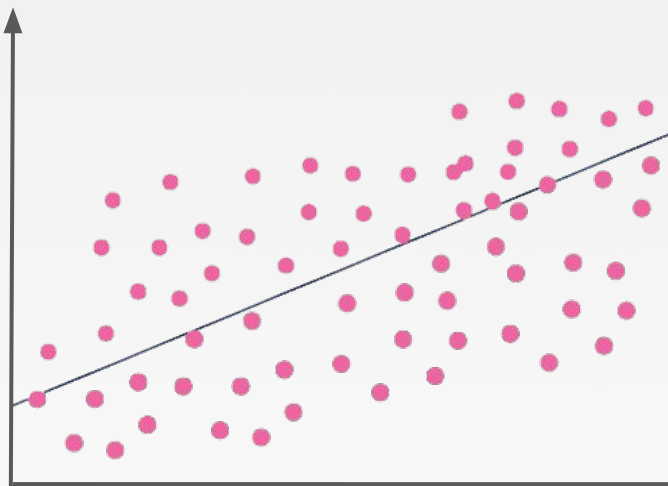


**GHOST**

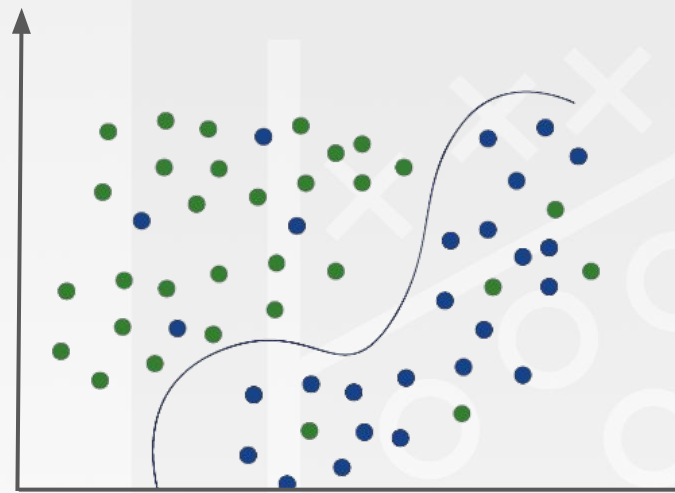
Group of Horribly Optimistic Statisticians



# Regresja vs klasyfikacja



Regresja



Klasyfikacja



**GHOST**

Group of Horribly Optimistic Statisticians

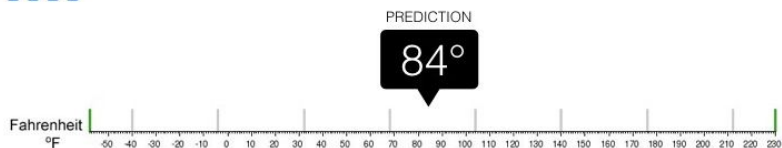


# Regresja vs klasyfikacja



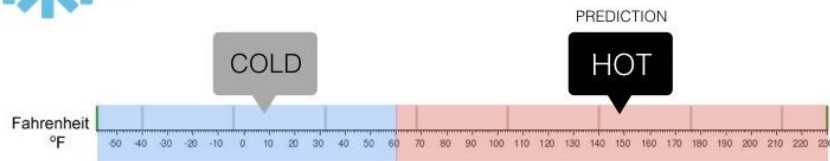
## Regression

What is the temperature going to be tomorrow?



## Classification

Will it be Cold or Hot tomorrow?





**GHOST**

Group of Horribly Optimistic Statisticians

# Klasyfikacja podstawowe pojęcia



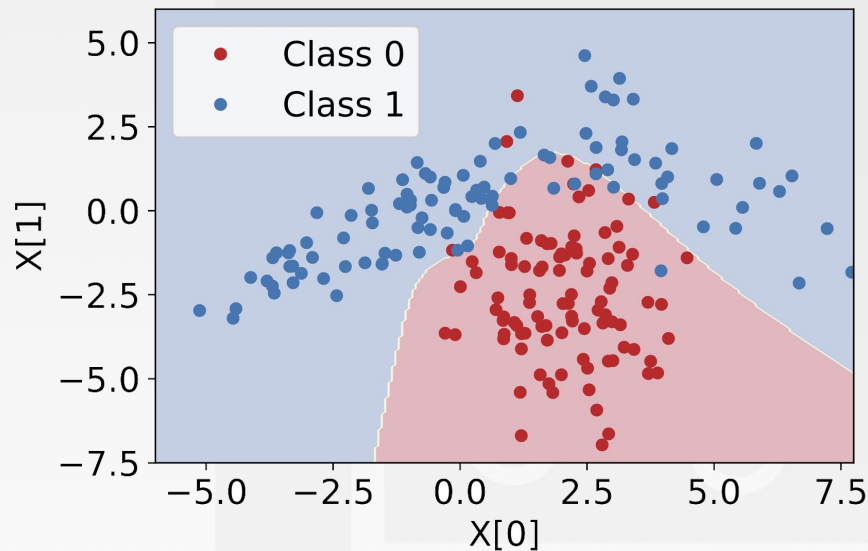
**GHOST**

Group of Horribly Optimistic Statisticians



# Czym jest klasyfikacja?

Klasyfikacja to proces, w którym na podstawie dostępnych danych przewiduje się wartość określonego atrybutu. Jej celem jest przypisanie danego obiektu do jednej z wcześniej ustalonych kategorii na podstawie jego cech.



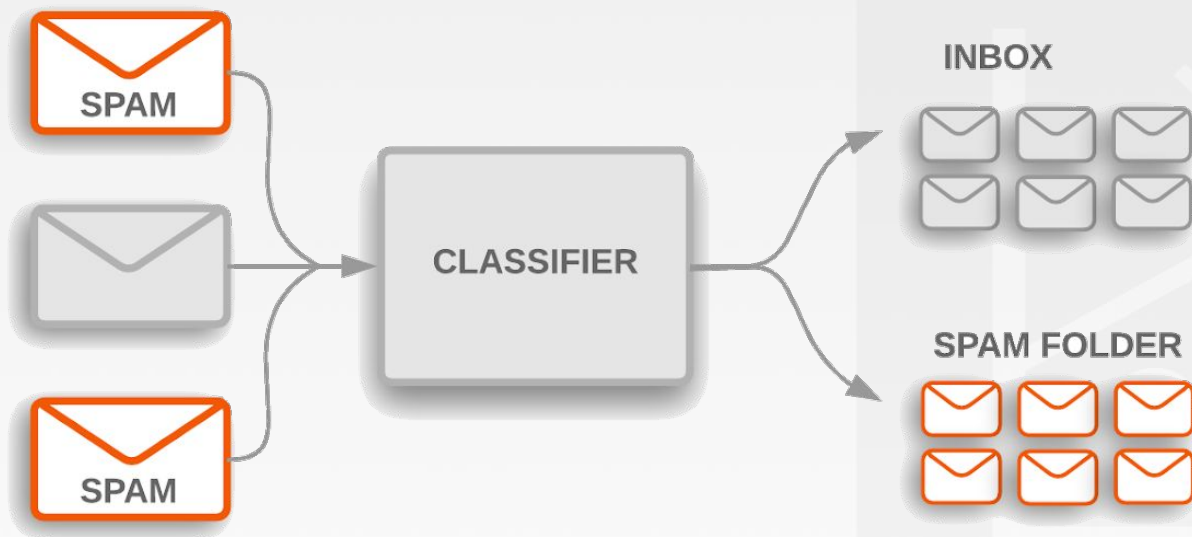


**GHOST**

Group of Horribly Optimistic Statisticians



## Zastosowanie - filtrowaniu spamu





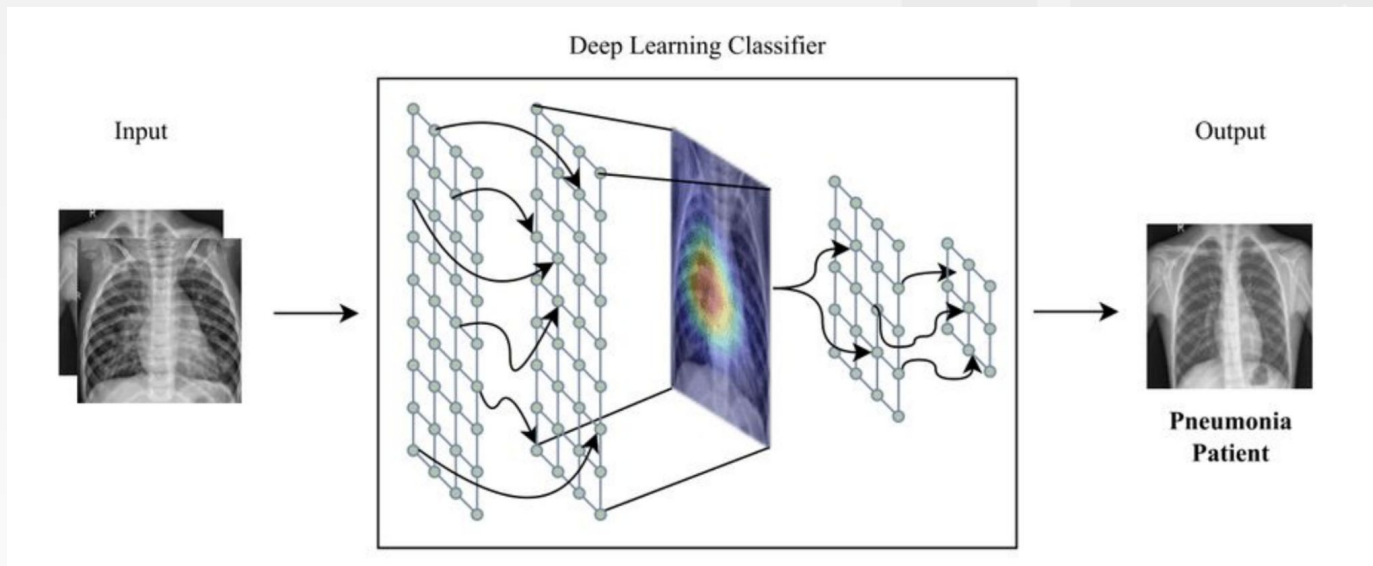


# GHOST

Group of Horribly Optimistic Statisticians



Zastosowanie - **diagnozowanie chorób** (na podstawie danych medycznych, takich jak wyniki badań, objawy oraz historia choroby).





# GHOST

Group of Horribly Optimistic Statisticians



Zastosowanie - **wykrywanie nadużyć finansowych** ( wykrywanie działania o charakterze przestępczym poprzez analizę wzorców transakcji i identyfikację anomalii)



The traditional approach identify fraudulent activities through known past behaviour



The machine learning approach models a user banking patterns and detect anomalous behaviours

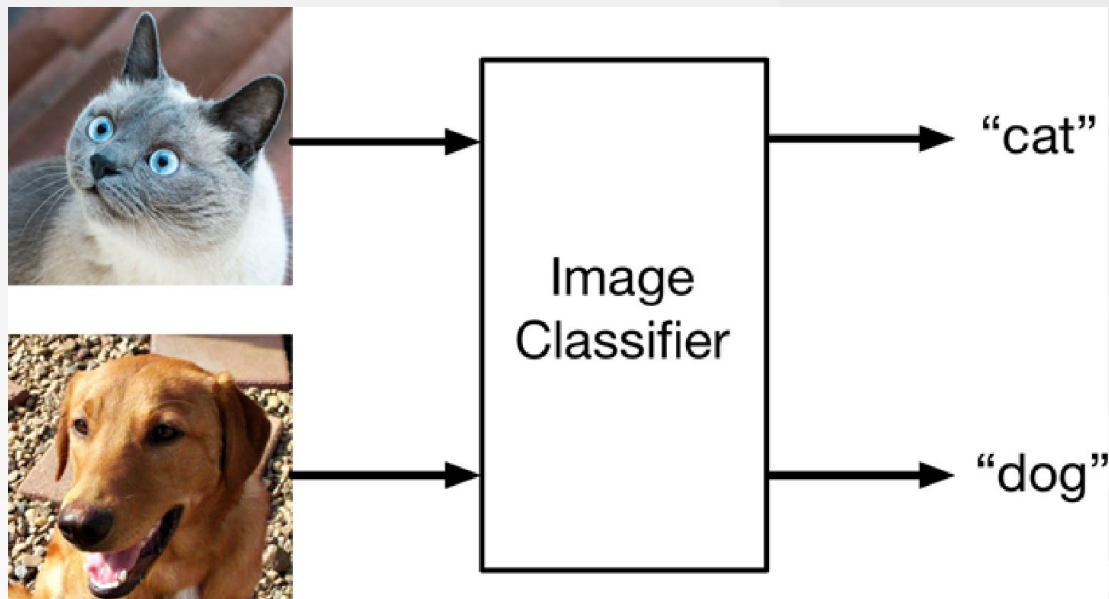


**GHOST**

Group of Horribly Optimistic Statisticians



## Zastosowanie - klasyfikacja obrazów





**GHOST**

Group of Horribly Optimistic Statisticians



# Ważne zagadnienia związane z klasyfikacją

## 1. Cechy i etykiety

- Cechy(ang. features)  
- wejście
- Etykiety (ang. labels)  
- wyjście

Features		Label
size	edge	color
small	dotted	green
big	striped	yellow
medium	normal	green



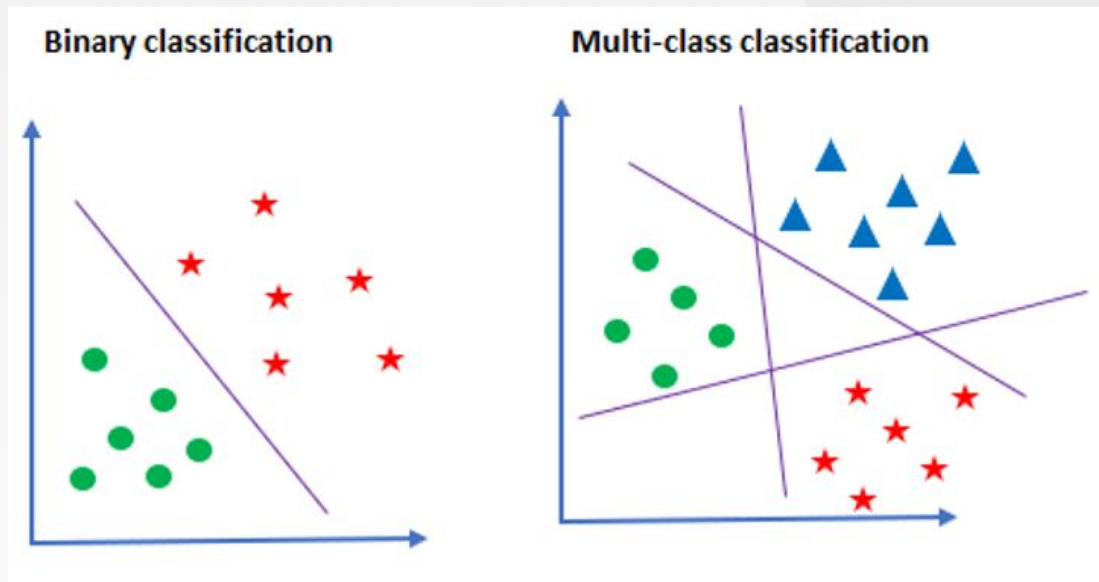
**GHOST**

Group of Horribly Optimistic Statisticians



# Ważne zagadnienia związane z klasyfikacją

## 1. Klasyfikacja binarna vs wieloklasowa







**GHOST**

Group of Horribly Optimistic Statisticians



# Ważne zagadnienia związane z klasyfikacją

## 1. Klasyfikacja wieloklasowa vs wieloetykietowa

	Multi-Class	Multi-Label
Image		
Labels	<ul style="list-style-type: none"><li>cat</li><li>✓ dog</li><li>hamster</li><li>rabbit</li><li>fish</li></ul>	<ul style="list-style-type: none"><li>✓ dog</li><li>long-haired</li><li>✓ glasses</li><li>ears up</li><li>✓ collar</li></ul>



**GHOST**

Group of Horribly Optimistic Statisticians

# Metody oceny klasyfikacji



# GHOST

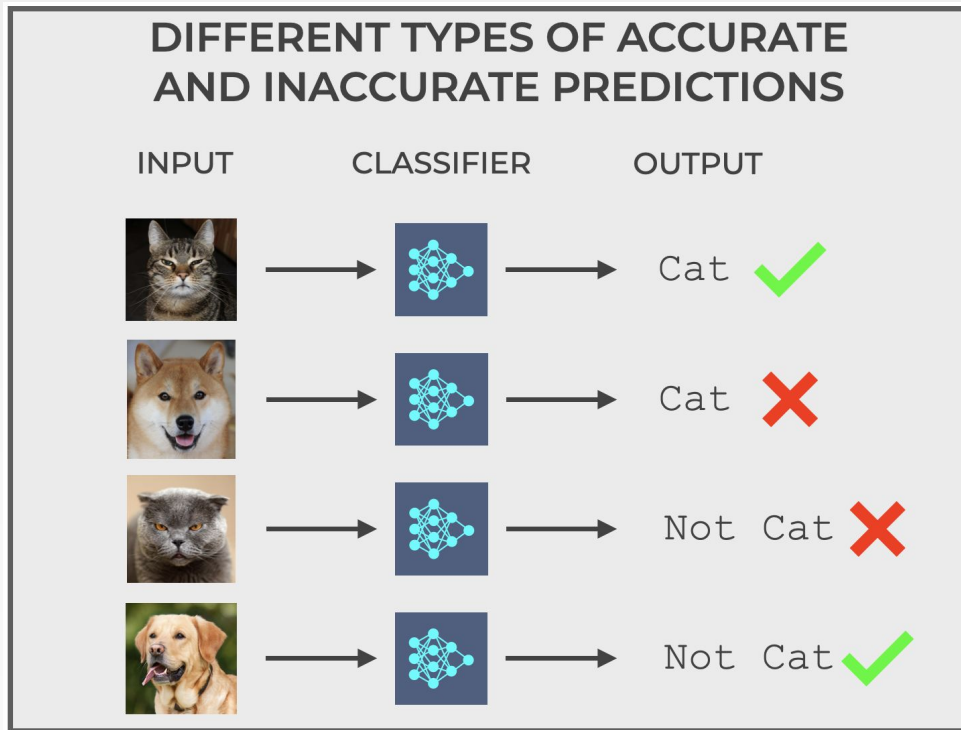
Group of Horribly Optimistic Statisticians



Czy klasyfikator  
popętnia błędy?

Jak często się one  
zdarzają?

Jak ocenić czy pomimo  
pomyłki resztę dobrze  
przypasował?







# GHOST

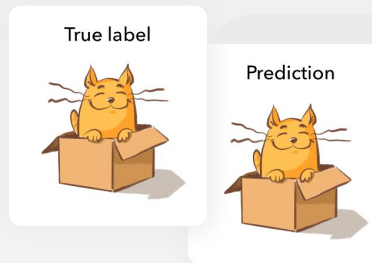
Group of Horribly Optimistic Statisticians



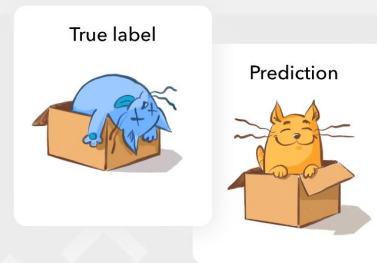
## Macierz pomyłek (eng. Confusion matrix)

		Wartość prognozowana	
		Negatywna	Pozytywna
Wartość rzeczywista	Negatywna	Prawdziwie Negatywna (TN)	Fałszywie Pozytywna (FP)
	Pozytywna	Fałszywie Negatywna (FN)	Prawdziwie Pozytywna (TP)

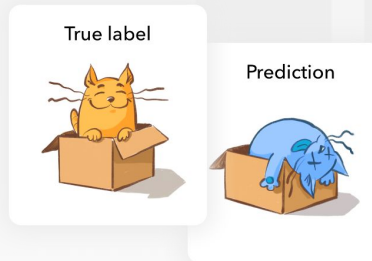
✓ TRUE POSITIVE



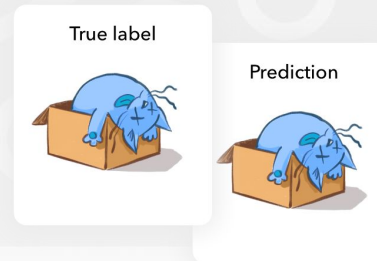
✗ FALSE POSITIVE



✗ FALSE NEGATIVE



✓ TRUE NEGATIVE



**GHOST**

Group of Horribly Optimistic Statisticians



Na podstawie macierzy pomyłek oblicza się różne miary skuteczności:

- **Dokładność (Accuracy)** określa, jaki procent wszystkich przewidywań modelu był poprawny – zarówno pozytywnych, jak i negatywnych

		Wartość prognozowana	
		Negatywna	Pozytywna
Wartość rzeczywista	Negatywna	Prawdziwie Negatywna (TN)	Fałszywie Pozytywna (FP)
	Pozytywna	Fałszywie Negatywna (FN)	Prawdziwie Pozytywna (TP)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$



# GHOST

Group of Horribly Optimistic Statisticians



Na podstawie macierzy pomyłek oblicza się różne miary skuteczności:

→ **Precyzja (Precision)** – odsetek prawdziwie pozytywnych wyników spośród wszystkich przypadków zaklasyfikowanych jako pozytywne

		Wartość prognozowana	
		Negatywna	Pozytywna
Wartość rzeczywista	Negatywna	Prawdziwie Negatywna (TN)	Fałszywie Pozytywna (FP)
	Pozytywna	Fałszywie Negatywna (FN)	Prawdziwie Pozytywna (TP)

$$\textbf{Precision} = \frac{TP}{TP + FP}$$



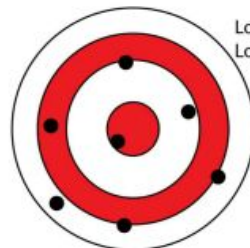
# GHOST

Group of Horribly Optimistic Statisticians



Accuracy vs precision

## Accuracy vs Precision



Low accuracy  
Low precision



Low accuracy  
High precision



High accuracy  
Low precision



High accuracy  
High precision



# GHOST

Group of Horribly Optimistic Statisticians



Na podstawie macierzy pomyłek oblicza się różne miary skuteczności:

→ **Czułość (Recall/Sensitivity)** – odsetek prawdziwie pozytywnych wyników spośród wszystkich rzeczywiście pozytywnych przypadków

		Wartość prognozowana	
		Negatywna	Pozytywna
Wartość rzeczywista	Negatywna	Prawdziwie Negatywna (TN)	Fałszywie Pozytywna (FP)
	Pozytywna	Fałszywie Negatywna (FN)	Prawdziwie Pozytywna (TP)

$$\text{Recall} = \frac{TP}{TP + FN}$$

**GHOST**

Group of Horribly Optimistic Statisticians



Na podstawie macierzy pomyłek oblicza się różne miary skuteczności:

- **Specyficzność (Specificity)** – odsetek prawdziwie negatywnych wyników spośród wszystkich rzeczywiście negatywnych przypadków

		Wartość prognozowana	
		Negatywna	Pozytywna
Wartość rzeczywista	Negatywna	Prawdziwie Negatywna (TN)	Fałszywie Pozytywna (FP)
	Pozytywna	Fałszywie Negatywna (FN)	Prawdziwie Pozytywna (TP)

$$\textit{Specificity} = \frac{TN}{TN + FP}$$



# GHOST

Group of Horribly Optimistic Statisticians



## Macierz pomyłek w praktyce:

	Predicted class POSITIVE (spam ✉ )	Predicted class NEGATIVE (normal 📧 )
Actual class POSITIVE (spam ✉ )	TRUE POSITIVE (TP) ✉ ✉ 320	FALSE NEGATIVE (FN) ✉ 📧 43
Actual class NEGATIVE (normal 📧 )	FALSE POSITIVE (FP) ✉ ✉ 20	TRUE NEGATIVE (TN) 📧 📧 538

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP} \\ &= \frac{320}{320 + 20} = 0.941 \end{aligned}$$

$$\begin{aligned} \text{Recall} &= \frac{TP}{TP + FN} \\ &= \frac{320}{320 + 43} = 0.882 \end{aligned}$$

$$\begin{aligned} \text{Specificity} &= \frac{TN}{FP + TN} \\ &= \frac{538}{20 + 538} = 0.964 \end{aligned}$$

Accuracy?



**GHOST**

Group of Horribly Optimistic Statisticians

# Proste klasyfikatory





**GHOST**

Group of Horribly Optimistic Statisticians



## Regresja logistyczna

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n)}}$$



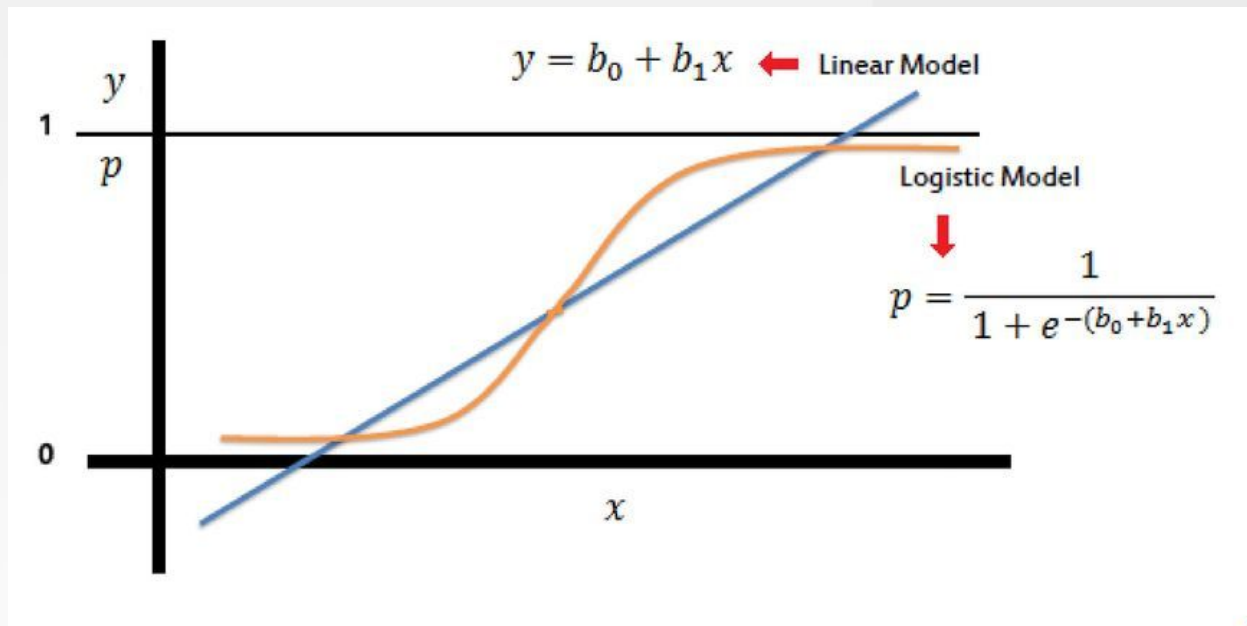


**GHOST**

Group of Horribly Optimistic Statisticians



## Regresja logistyczna vs Regresja liniowa



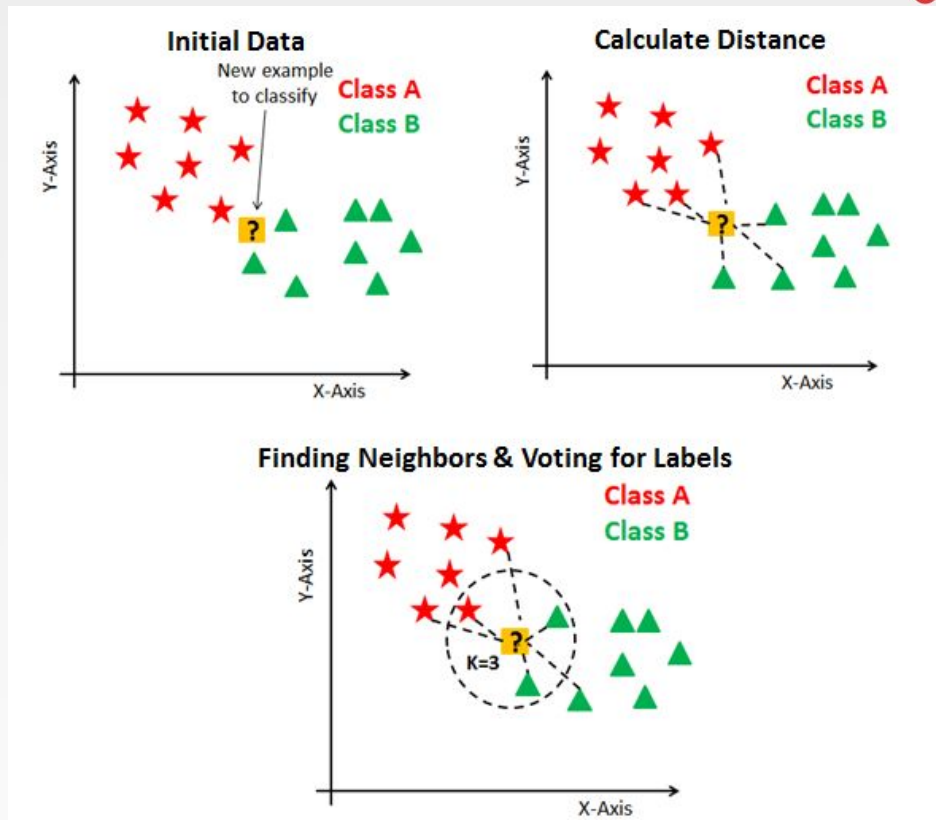


**GHOST**

Group of Horribly Optimistic Statisticians

## K najbliższych sąsiadów (KNN)

k – parametr określający k  
najbliższych sąsiadów





**GHOST**

Group of Horribly Optimistic Statisticians

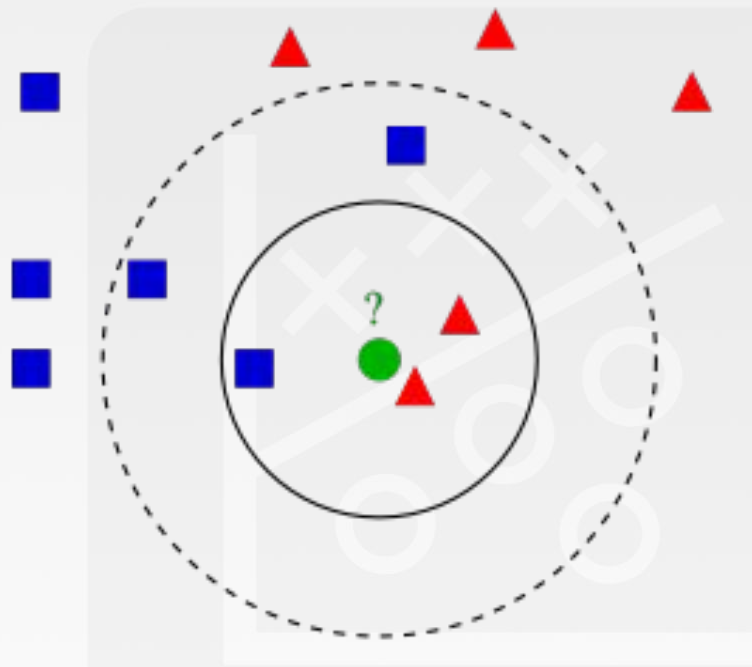


## K najbliższych sąsiadów (KNN)

Przewidywana klasa dla zielonego  
kółka jeśli:

- $k = 1$
- $k = 3$
- $k = 5$

to ...





**GHOST**

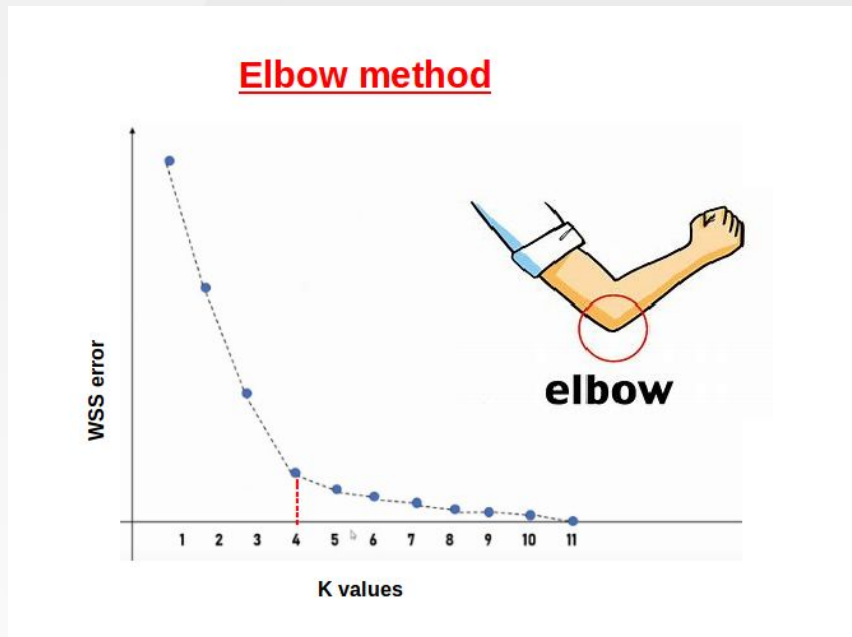
Group of Horribly Optimistic Statisticians



## K najbliższych sąsiadów (KNN)

Jak najlepiej dopasować parametr  $k$  ?

- Walidacja Krzyżowa  
(ang. Cross Validation)
- Metoda "łokcia"  
(ang. Elbow Method)
- Nieparzyste wartości  $k$



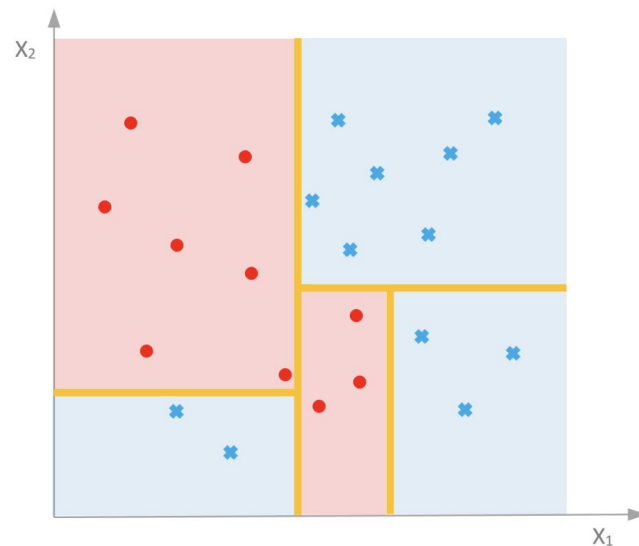
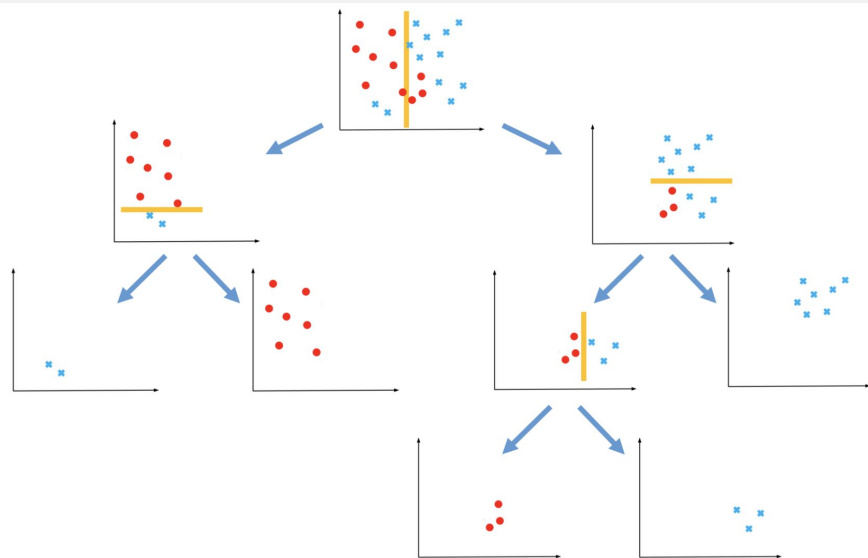


**GHOST**

Group of Horribly Optimistic Statisticians



## Drzewo decyzyjne - idea





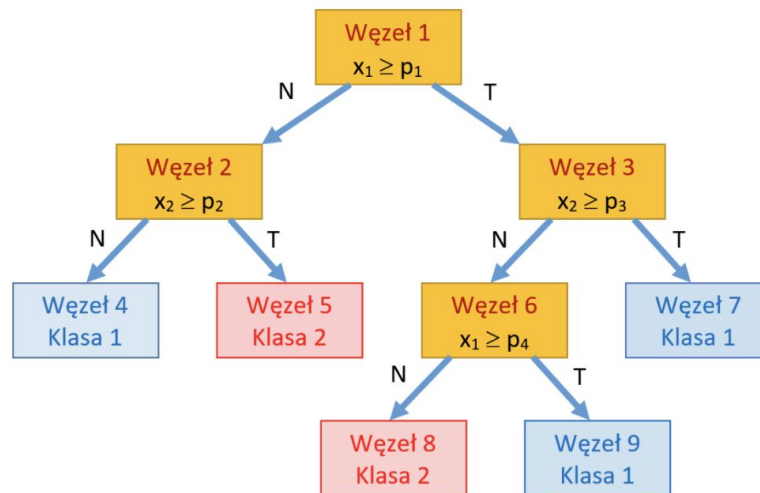
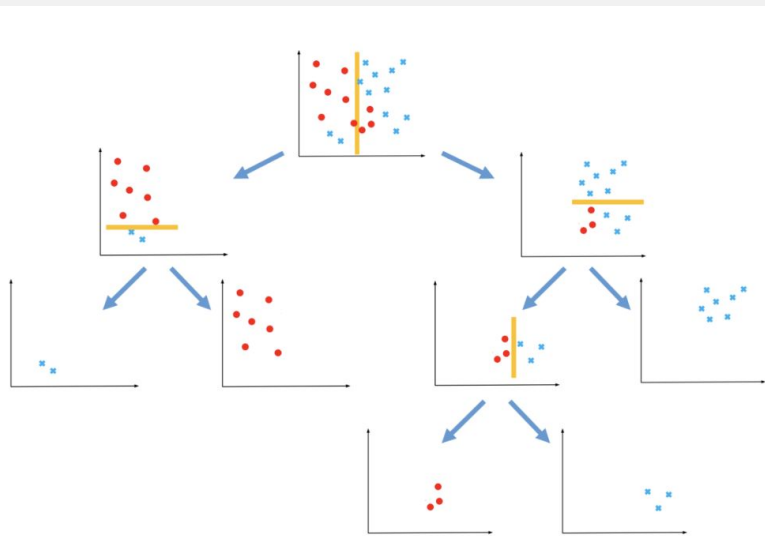
**GHOST**

Group of Horribly Optimistic Statisticians



# Drzewo decyzyjne - konstruowanie

Cel: Maksymalne różnicowanie klas w podzbiorach



**GHOST**

Group of Horribly Optimistic Statisticians

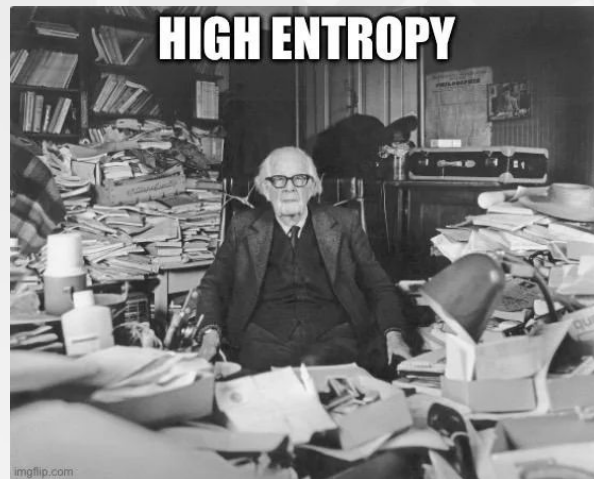


## Drzewo decyzyjne - Jak ocenić różnicowanie klas?

Jako miary rozkładu klas w zbiorze przykładów (miary zanieczyszczenia węzła, miary informacji w zbiorze) używa się:

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

**p** - prawdopodobieństwo otrzymania  
wybranej klasy w zbiorze





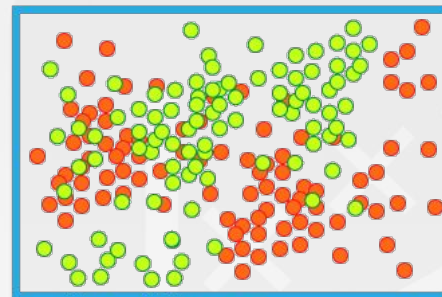
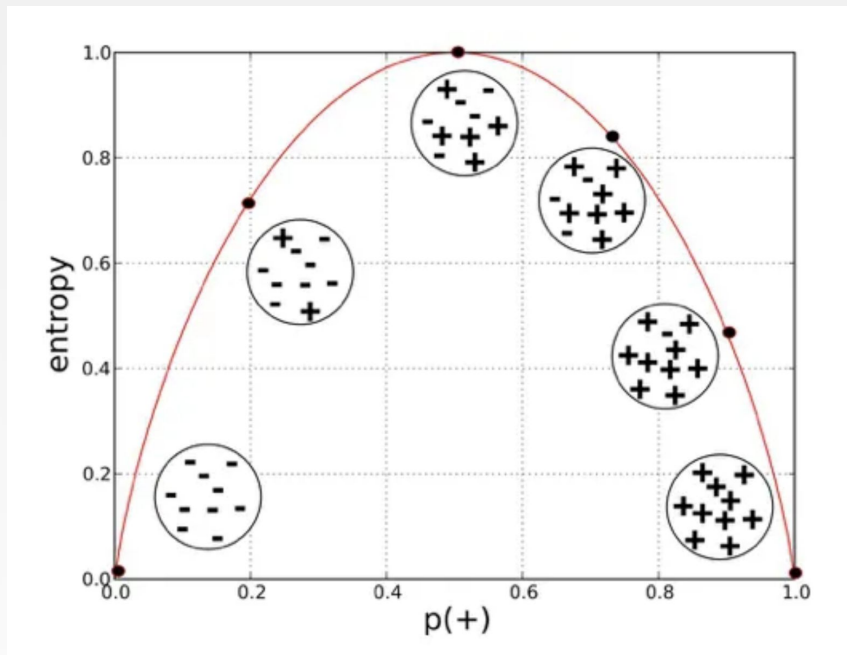


**GHOST**

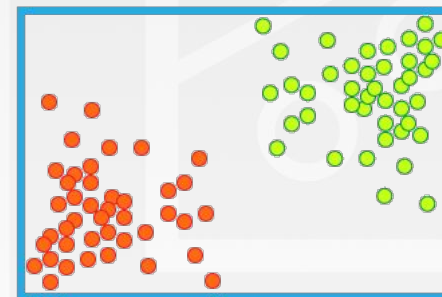
Group of Horribly Optimistic Statisticians



# Drzewo decyzyjne - Entropia jako nasz przyjaciel



High Entropy



Low Entropy



# GHOST

Group of Horribly Optimistic Statisticians



## Drzewo decyzyjne - przykład

$$E(\text{Parent}) = -\frac{16}{30}\log_2\left(\frac{16}{30}\right) - \frac{14}{30}\log_2\left(\frac{14}{30}\right) \approx 0.99$$

$$E(\text{Balance} < 50K) = -\frac{12}{13}\log_2\left(\frac{12}{13}\right) - \frac{1}{13}\log_2\left(\frac{1}{13}\right) \approx 0.39$$

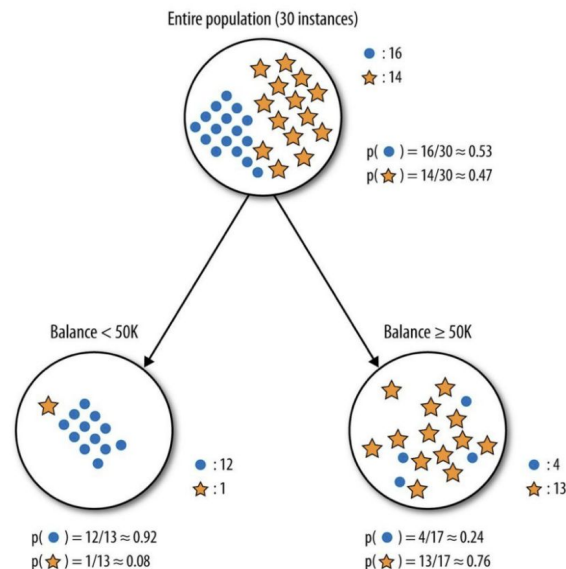
$$E(\text{Balance} > 50K) = -\frac{4}{17}\log_2\left(\frac{4}{17}\right) - \frac{13}{17}\log_2\left(\frac{13}{17}\right) \approx 0.79$$

Weighted Average of entropy for each node:

$$\begin{aligned} E(\text{Balance}) &= \frac{13}{30} \times 0.39 + \frac{17}{30} \times 0.79 \\ &= 0.62 \end{aligned}$$

Information Gain:

$$\begin{aligned} IG(\text{Parent}, \text{Balance}) &= E(\text{Parent}) - E(\text{Balance}) \\ &= 0.99 - 0.62 \\ &= 0.37 \end{aligned}$$





# GHOST

Group of Horribly Optimistic Statisticians



## Drzewo decyzyjne - przykład

$$E(\text{Residence} = \text{OWN}) = -\frac{7}{8}\log_2\left(\frac{7}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right) \approx 0.54$$

$$E(\text{Residence} = \text{RENT}) = -\frac{4}{10}\log_2\left(\frac{4}{10}\right) - \frac{6}{10}\log_2\left(\frac{6}{10}\right) \approx 0.97$$

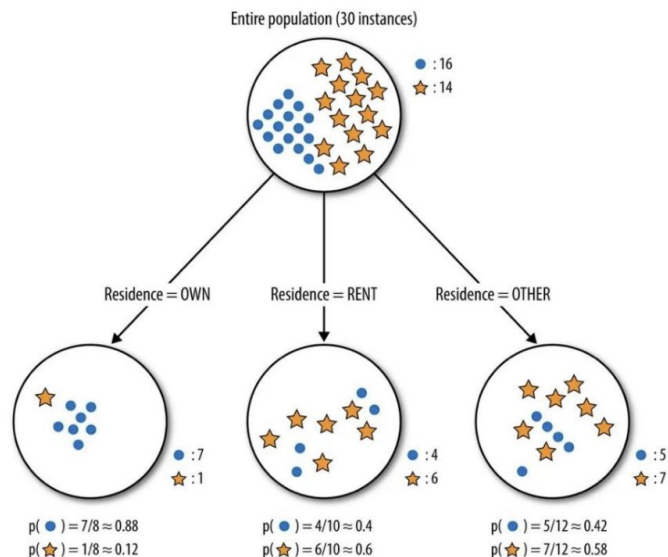
$$E(\text{Residence} = \text{OTHER}) = -\frac{5}{12}\log_2\left(\frac{5}{12}\right) - \frac{7}{12}\log_2\left(\frac{7}{12}\right) \approx 0.98$$

Weighted Average of entropies for each node:

$$E(\text{Residence}) = \frac{8}{30} \times 0.54 + \frac{10}{30} \times 0.97 + \frac{12}{30} \times 0.98 = 0.86$$

Information Gain:

$$\begin{aligned} IG(\text{Parent}, \text{Residence}) &= E(\text{Parent}) - E(\text{Residence}) \\ &= 0.99 - 0.86 \\ &= 0.13 \end{aligned}$$





**GHOST**

Group of Horribly Optimistic Statisticians

# Inne wyzwania



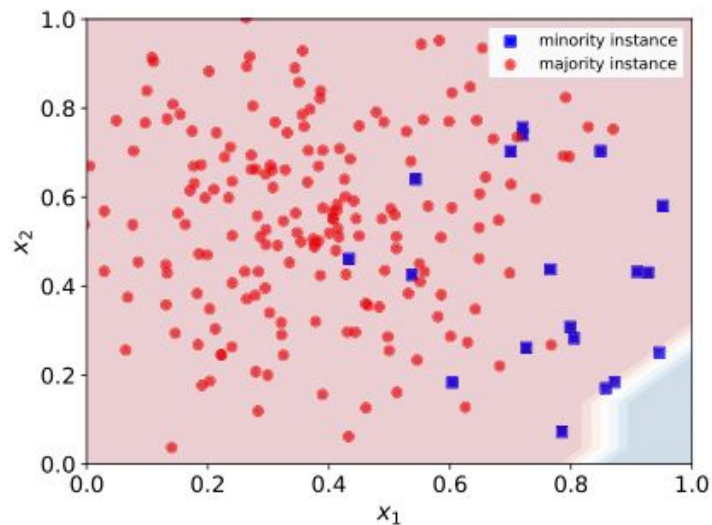
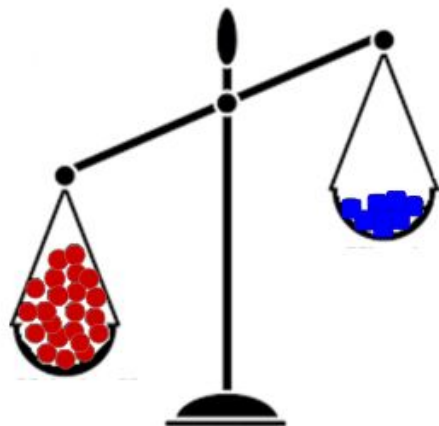


**GHOST**

Group of Horribly Optimistic Statisticians



# Niezbalansowanie klas





**GHOST**

Group of Horribly Optimistic Statisticians

# Niezbalansowanie klas - metryki

## Weighted Balanced Accuracy

$$WeightedBalancedAccuracy = \sum_{i=1}^C w_i * Accuracy_i$$

$$0 \leq w_i \leq 1, \text{ and, } \sum_{i=1}^C w_i = 1$$



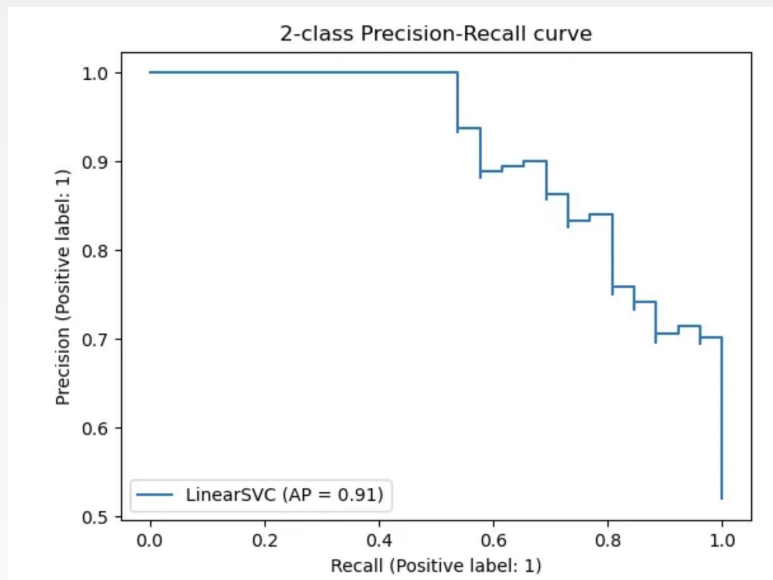


**GHOST**

Group of Horribly Optimistic Statisticians



# Niezbalansowanie klas - metryki Precision-Recall Curve(AUC-PR)



Wysoka wartość AUC-PR (Area Under Curve)

- zwraca **trafne przewidywania** (wysoka precyzja),
- **wychwytuje większość pozytywnych przypadków** (wysoka czułość).

**Wysoka precyzja** → mało fałszywie pozytywnych wyników.

**Wysoka czułość** → mało fałszywie negatywnych wyników.

<https://arize.com/blog/what-is-pr-auc/>



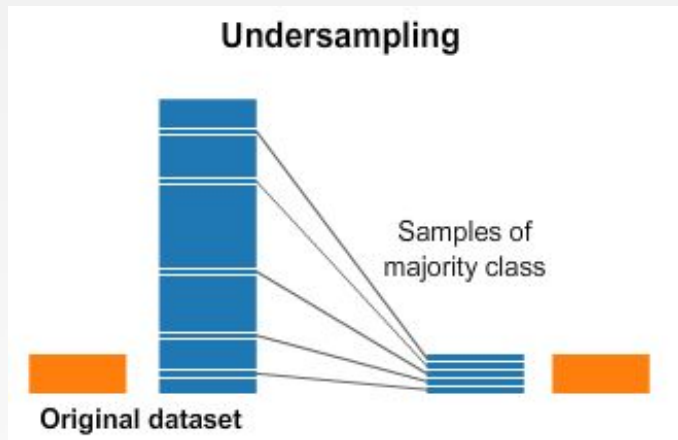
**GHOST**

Group of Horribly Optimistic Statisticians

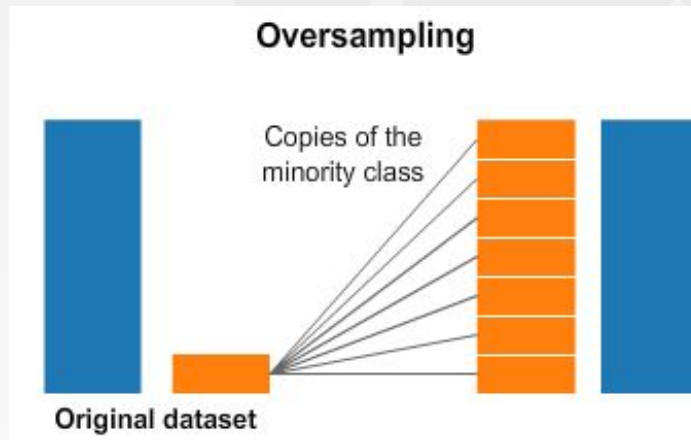


# Niezbalansowanie klas - inne

Podpróbkiwanie



Nadpróbkiwanie







**GHOST**

Group of Horribly Optimistic Statisticians



## Klątwa wymiarowości



**VS**



ogniste



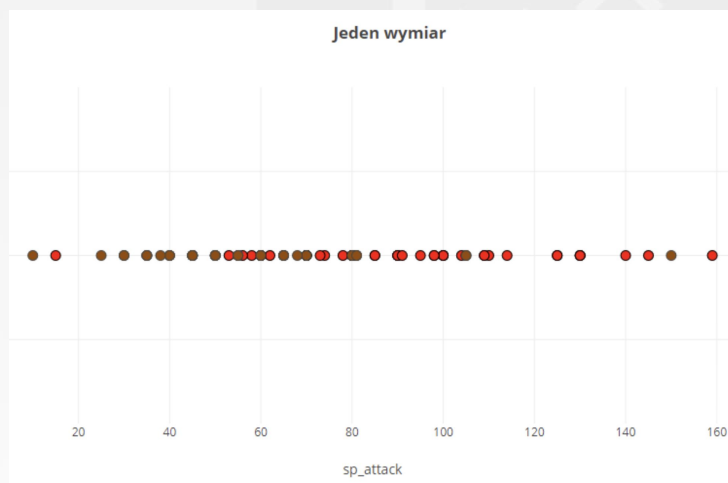
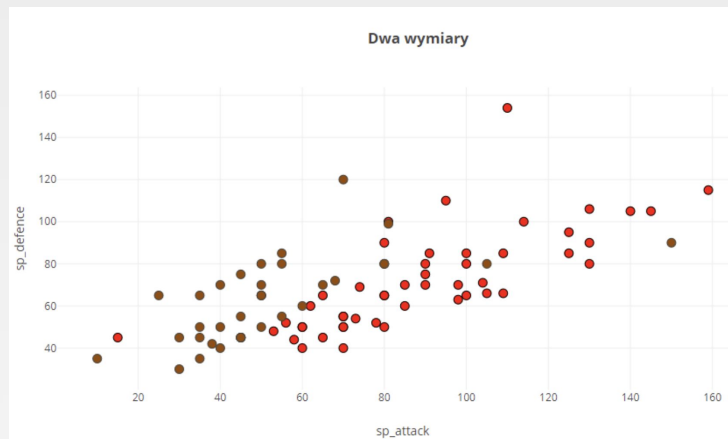
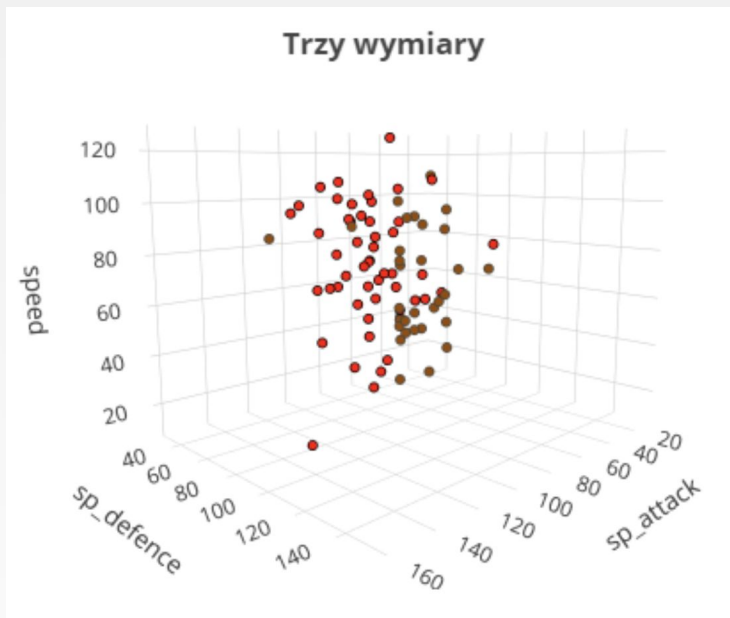
ziemne



# GHOST

Group of Horribly Optimistic Statisticians

## Klątwa wymiarowości



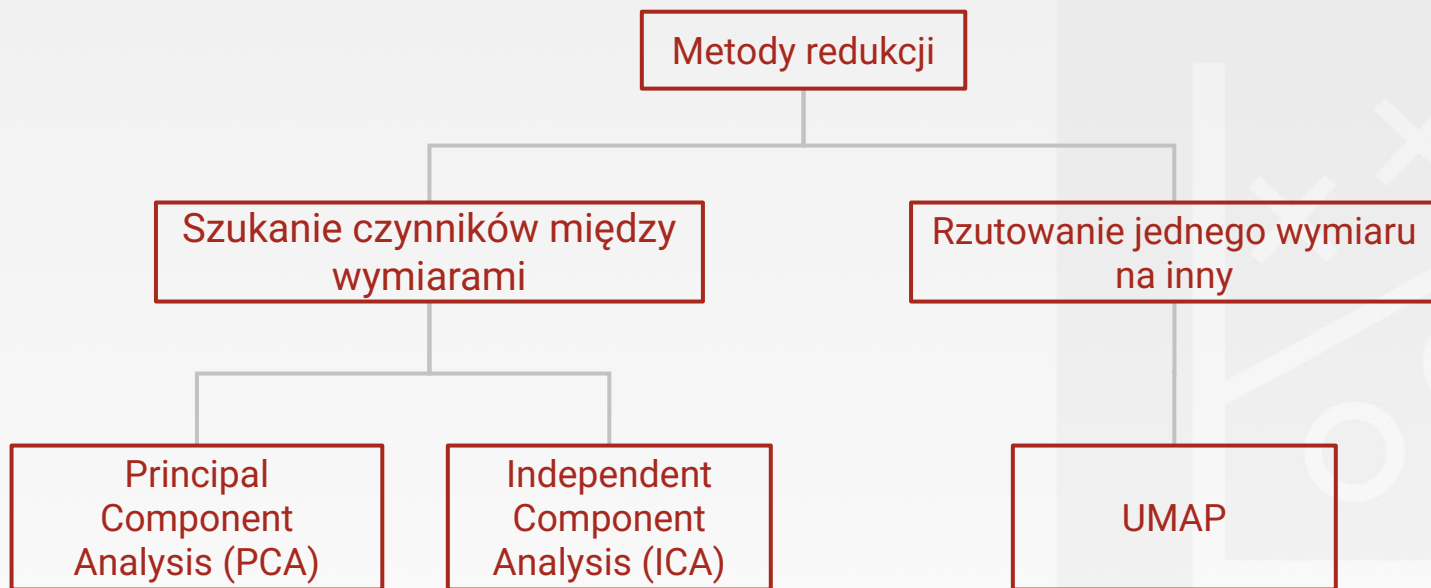


**GHOST**

Group of Horribly Optimistic Statisticians



# Klątwa wymiarowości - czy 3 cechy mogą zastąpić nam 100?



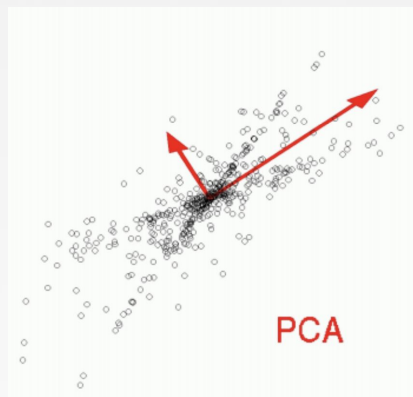


**GHOST**

Group of Horribly Optimistic Statisticians

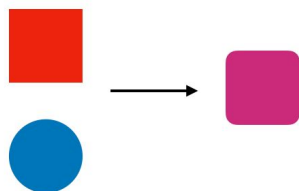


# Klątwa wymiarowości - PCA vs ICA



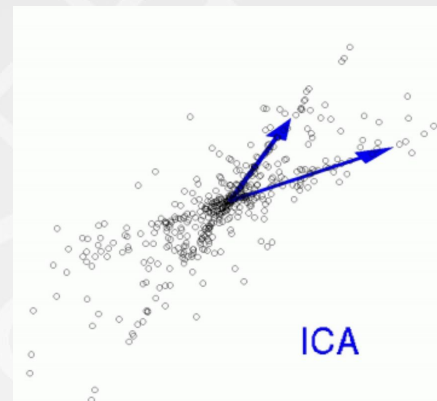
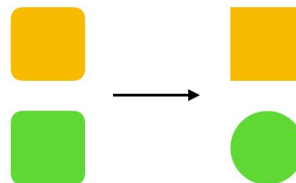
**PCA**

**Compresses information**



**ICA**

**Separates information**





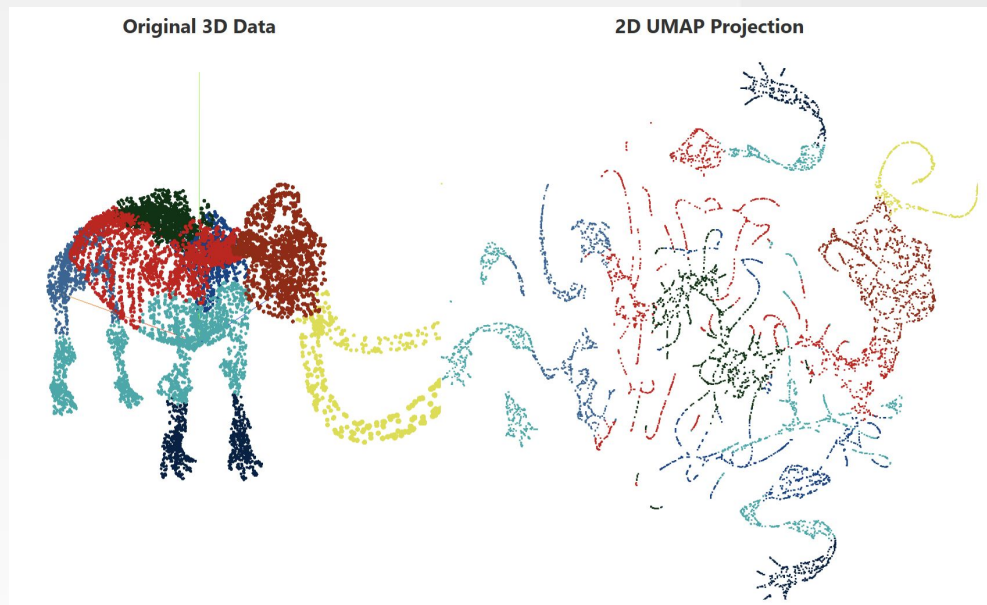
**GHOST**

Group of Horribly Optimistic Statisticians



# Klątwa wymiarowości - UMAP

<https://pair-code.github.io/understanding-umap/>

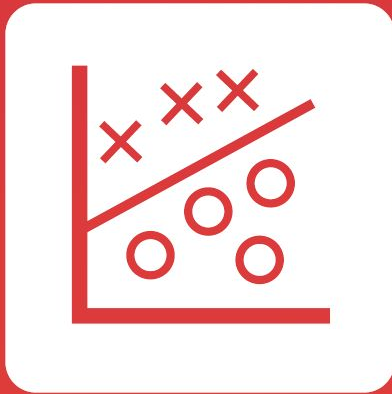




**GHOST**

Group of Horribly Optimistic Statisticians

Dziękuję za uwagę!



# GHOST

Group of Horribly Optimistic Statisticians