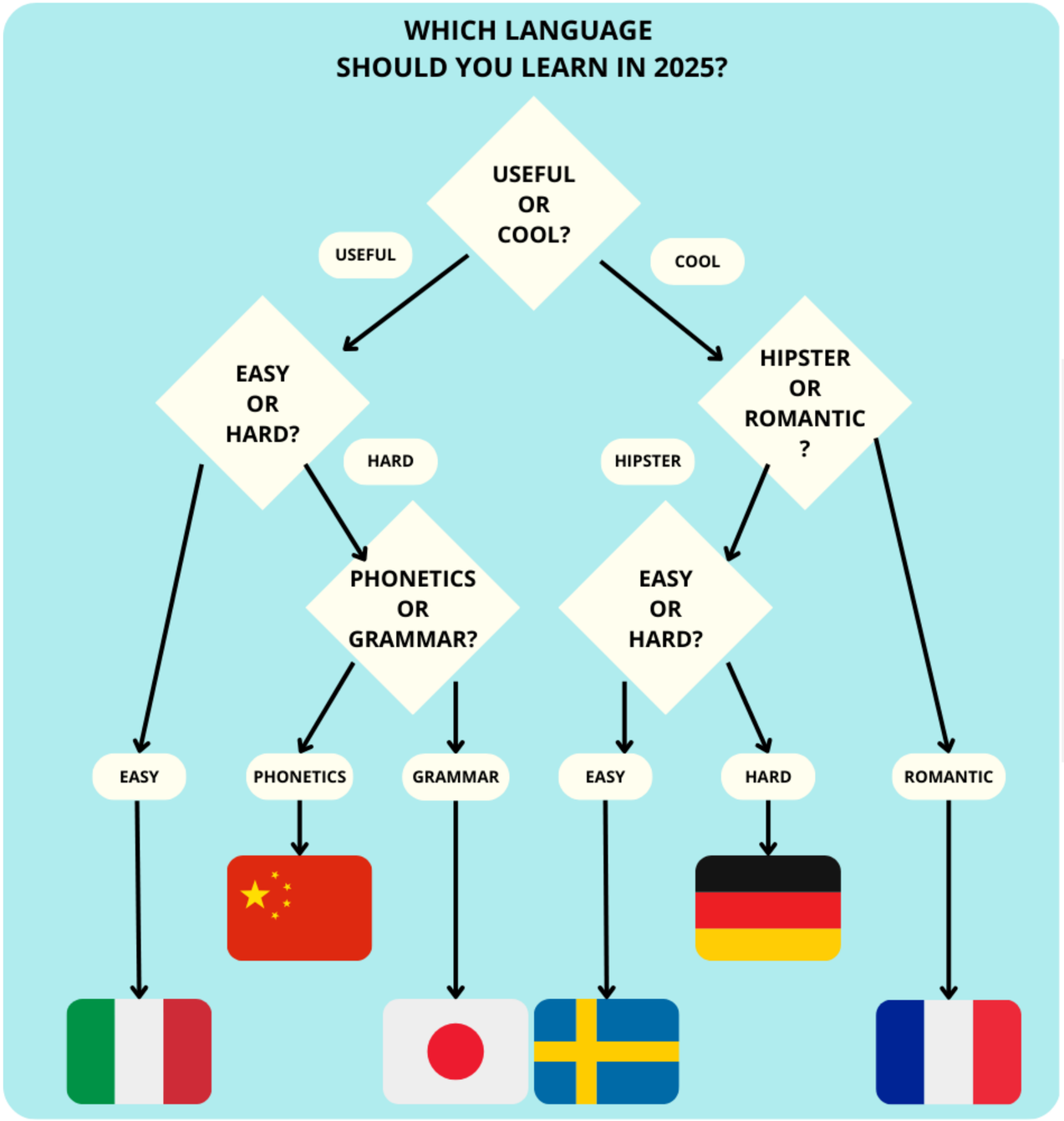
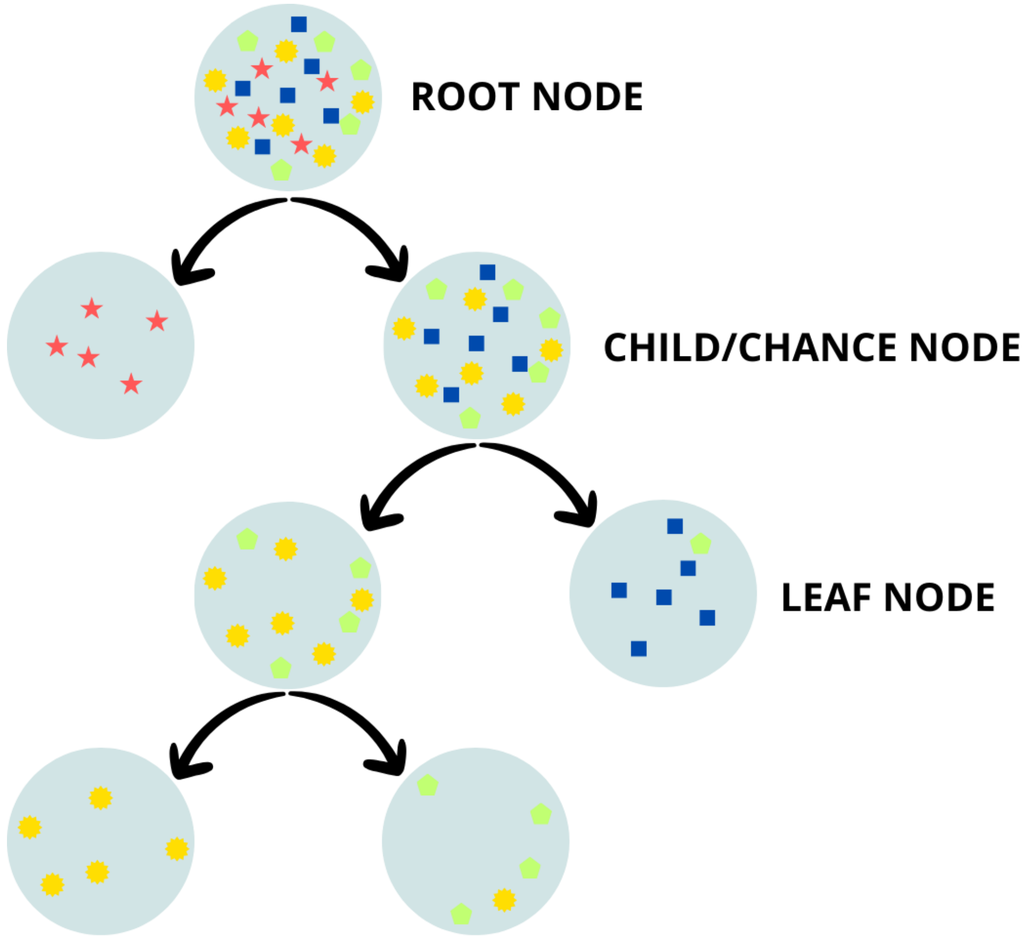
Introduction

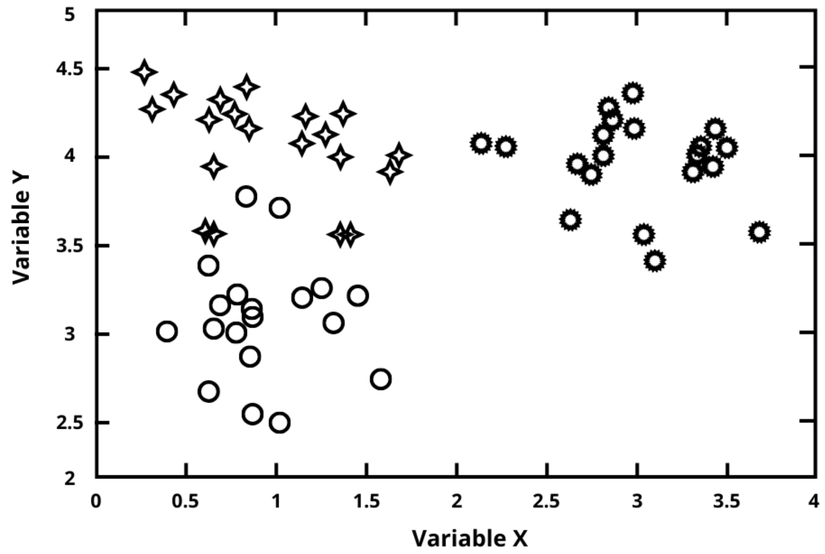


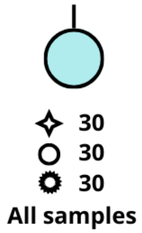
Decision Tree



Drzewo decyzyjne to prosty model stosowany w nadzorowanej klasyfikacji. Wykorzystuje się je do klasyfikowania pojedynczej, dyskretnej cechy docelowej. Każdy węzeł wewnętrzny wykonuje test logiczny (Boolean) na jednej z cech wejściowych (ogólnie test może mieć więcej niż dwie opcje, ale można je przekształcić w szereg testów logicznych). Krawędzie są oznaczone wartościami tej cechy wejściowej. Każdy węzeł liściowy określa wartość cechy docelowej.

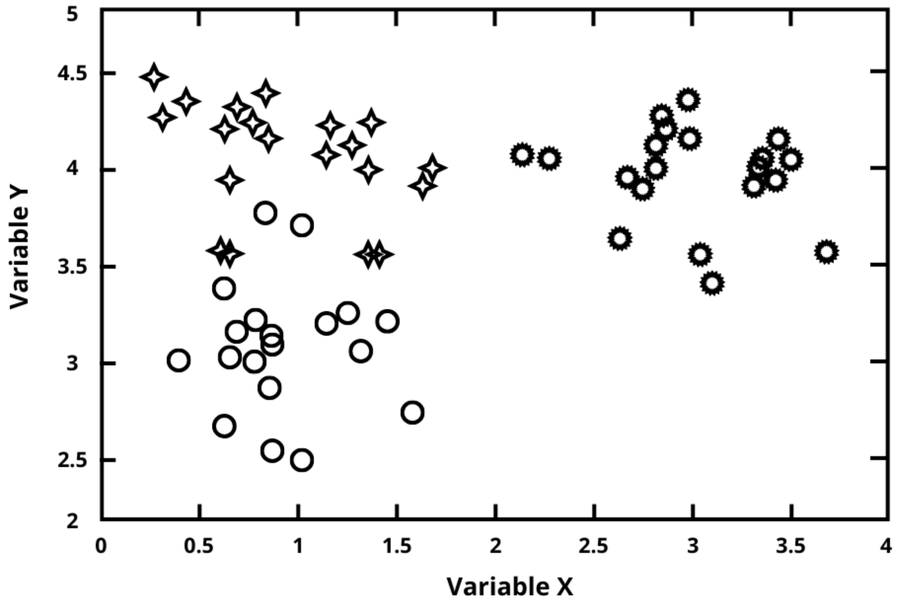
How to build a Decision Tree?

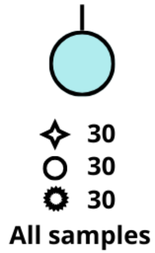


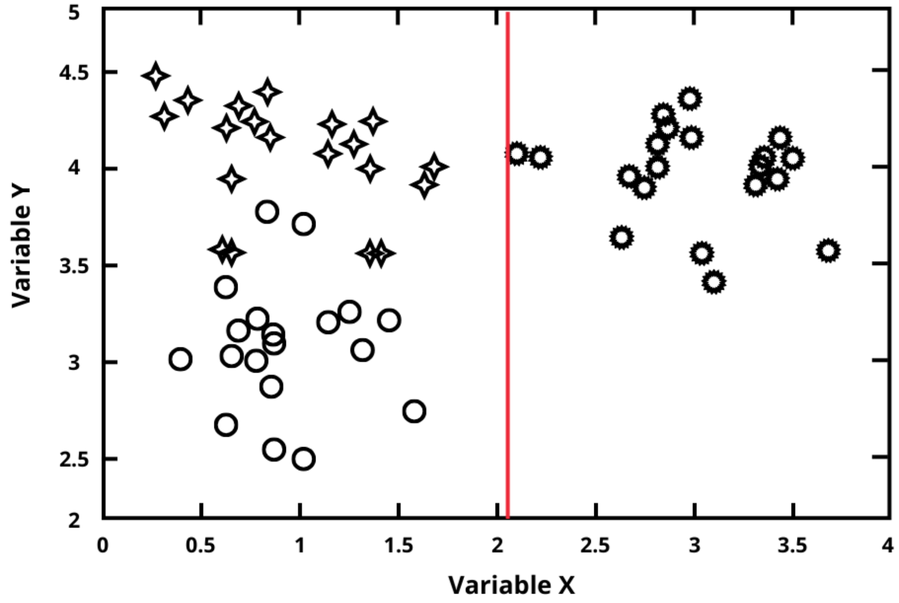


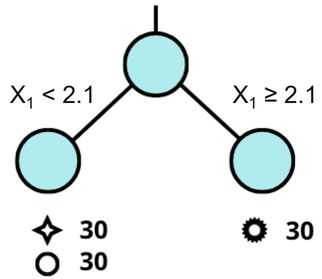
Source: Stanimirova I., et al. (2008) Chemometria w analityce.

How to build a Decision Tree?





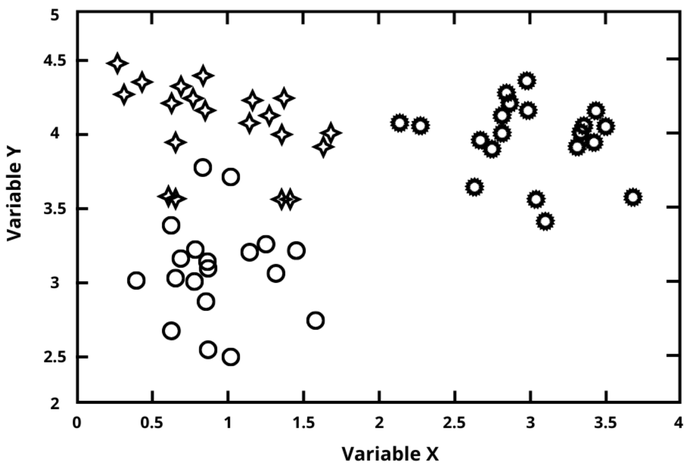


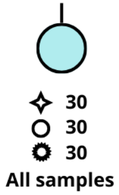


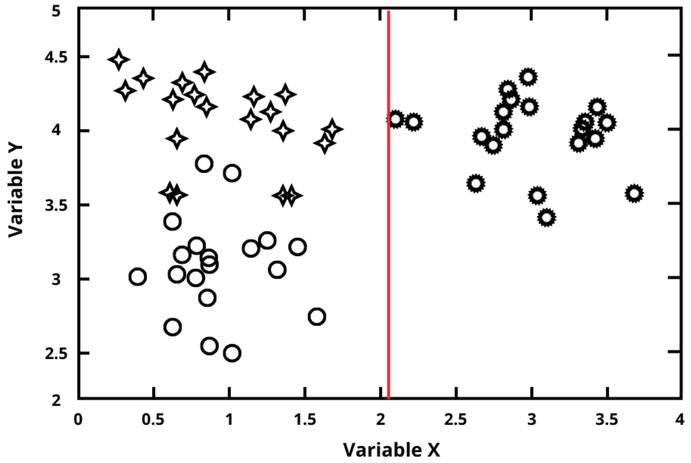
Construction of the 1st logical rule

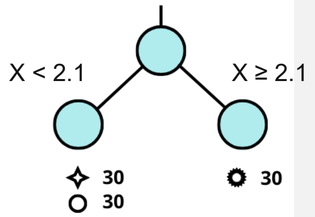
Source: Stanimirova I., et al. (2008) Chemometria w analityce.

How to build a Decision Tree?

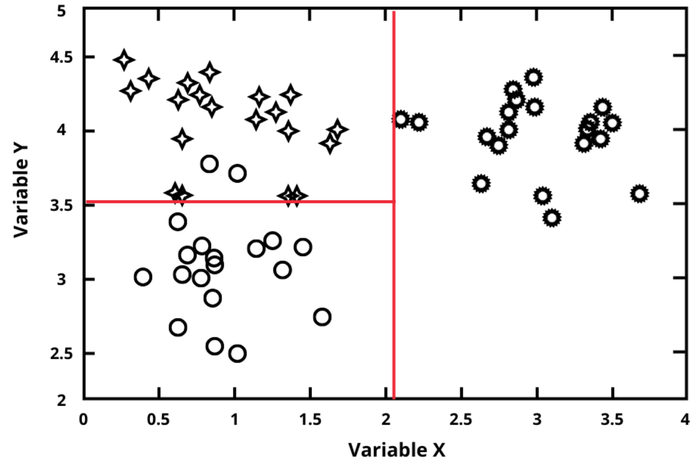


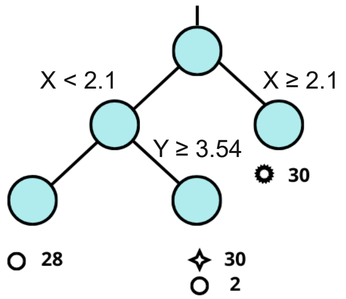






Construction of the 1st logical rule





Construction of the 2nd logical rule

Uniwersytet Gdański

Source: Stanimirova I., et al. (2008) Chemometria w analityce.

Reduction of node impurities

I. Gini value/index:

$$GINI(L) = 1 - \sum_{i=1}^{j} p_i^2$$

Gdzie: L oznacza zbiór obiektów, j to etykiety określające przynależność do klasy, a pi to przypisane prawdopodobieństwa dla grupy i w zbiorze L.

II. Entropy:

$$I(t) = -\sum_{i=1}^{k} p_i(t)\ln(p_i(t))$$

Gdzie: k oznacza liczbę grup próbek, a pi to odsetek próbek z i-tej grupy w węźle t.

GINI:

W algorytmach drzew decyzyjnych indeks Giniego jest miarą nieczystości lub nieuporządkowania w zbiorze danych. Pomaga określić najlepsze podziały, kwantyfikując, jak bardzo prawdopodobne jest, że dwa losowo wybrane elementy ze zbioru danych będą należeć do różnych klas.
Indeks Giniego jest cennym narzędziem w algorytmach drzew decyzyjnych, ponieważ pomaga wybrać najlepsze podziały i budować dokładne modele klasyfikacyjne. Minimalizując nieczystość Giniego, model dąży do tworzenia możliwie najczystszych węzłów, co poprawia jego zdolność do prawidłowej klasyfikacji nowych punktów danych.

ENTROPIA:

W algorytmach drzew decyzyjnych entropia jest kluczowym pojęciem służącym do mierzenia nieczystości lub nieuporządkowania w zbiorze danych. Dostarcza matematycznych podstaw do określania najlepszych podziałów w każdym węźle drzewa. Entropia odgrywa fundamentalną rolę w algorytmach drzew decyzyjnych, ponieważ kieruje wyborem podziałów, które maksymalizują przyrost informacji. Minimalizując entropię w każdym węźle, drzewo skutecznie dzieli dane na czystsze podzbiory, co prowadzi do lepszej jakości klasyfikacji.
W istocie, choć zarówno indeks Giniego, jak i entropia mierzą nieczystość, wykorzystują do tego nieco inne formuły matematyczne. Indeks Giniego skupia się na prawdopodobieństwie błędnej klasyfikacji losowo wybranego elementu, natomiast entropia opisuje niepewność lub losowość w zbiorze danych.

Confusion matrix

Prediction	Reality: Active	Reality: Inactive	
Active	**True positive** (TP)	**False positive** (FP) Type I Error	**Precision** $\frac{TP}{TP + FP}$
Inactive	**False negative** (FN) Type II Error	**True negative** (TN)	**Accuracy** $\frac{TP + TN}{TP + TN + FP + FN}$
Statistics	**Sensitivity** $\frac{TP}{TP + FN}$	**Specificity** $\frac{TN}{TN + FP}$	**F1** $\frac{2 \cdot Precision \cdot \text{Sensitivity}}{Precision + \text{Sensitivity}}$

Source: Gajewicz-Skretna A., et al. (2021) Ecotoxicol Environ Saf. 208:111738.

Macierz pomyłek (confusion matrix) jest podstawowym narzędziem służącym do oceny działania modelu klasyfikacyjnego, w tym drzew decyzyjnych. To tabela podsumowująca przewidywania modelu poprzez porównanie ich z rzeczywistymi etykietami (tzw. ground truth).
W macierzy pomyłek wyróżniamy cztery kluczowe pojęcia:

True Positives (TP) – przypadki, w których model poprawnie przewidział klasę pozytywną (np. prawidłowe wykrycie oszukańczej transakcji).
True Negatives (TN) – przypadki, w których model poprawnie przewidział klasę negatywną (np. poprawne sklasyfikowanie legalnej transakcji).
False Positives (FP) – przypadki, w których model błędnie przewidział klasę pozytywną, gdy rzeczywista klasa była negatywna (np. omyłkowe oznaczenie legalnej transakcji jako oszukańczej). To tzw. „błąd typu I".
False Negatives (FN) – przypadki, w których model błędnie przewidział klasę negatywną, gdy rzeczywista klasa była pozytywna (np. niewykrycie oszukańczej transakcji). To tzw. „błąd typu II".

Na podstawie tych wartości można obliczyć kilka podstawowych miar jakości modelu:
Accuracy (dokładność): mierzy ogólną poprawność modelu, czyli odsetek prawidłowych przewidywań względem wszystkich obserwacji.
Wzór: (TP + TN) / (TP + TN + FP + FN)
Precision (precyzja): określa odsetek poprawnych przewidywań klasy pozytywnej wśród wszystkich przewidywań pozytywnych.
Wzór: TP / (TP + FP)
Sensitivity (Recall, czułość): mierzy odsetek poprawnie wykrytych przypadków pozytywnych spośród wszystkich rzeczywistych przypadków pozytywnych.
Wzór: TP / (TP + FN)
Specificity (specyficzność): mierzy odsetek poprawnych przewidywań klasy negatywnej spośród wszystkich rzeczywistych przypadków negatywnych.
Wzór: TN / (TN + FP)
F1-score: średnia harmoniczna precyzji i czułości, jedna liczba podsumowująca obie miary.
Wzór: 2 * (Precision * Recall) / (Precision + Recall)

Analiza tych miar pozwala szczegółowo ocenić mocne i słabe strony modelu drzewa decyzyjnego, wskazać obszary wymagające poprawy oraz podjąć świadome decyzje dotyczące wdrażania i wykorzystania modelu w praktyce.

Confusion matrix

Balanced Accuracy (BAcc)	Balanced Error (BErr)	Matthews Correlation Coefficient
$\frac{\text{Sensitivity} + \text{Specificity}}{2}$	1 - BAcc	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Source: Gajewicz-Skretna A., et al. (2021) Ecotoxicol Environ Saf. 208:111738.

Zrównoważona dokładność (balanced accuracy) rozwiązuje kluczowy problem standardowej dokładności: jej podatność na niezrównoważone klasy. W przypadku zbiorów danych, w których jedna klasa znacząco dominuje nad drugą, samo maksymalizowanie ogólnej dokładności może być mylące. Model może osiągać wysoką dokładność, przewidując wyłącznie klasę większościową.

Zrównoważona dokładność eliminuje ten problem, obliczając średnią czułości (TPR) i specyficzności (TNR) dla każdej klasy. Dostarcza to bardziej wiarygodnej miary wydajności, zwłaszcza w sytuacjach z niezrównoważonymi danymi.

Z kolei *balanced error* bezpośrednio określa średni błąd dla obu klas. Oblicza się go jako **1 – balanced accuracy**. Niższy balanced error oznacza lepszą jakość modelu.

Współczynnik korelacji Matthewsa (MCC) jest bardziej ogólną miarą jakości klasyfikacji binarnych. Uwzględnia wszystkie cztery wartości z macierzy pomyłek (TP, TN, FP, FN) i daje wynik z zakresu od –1 do +1:

- **+1** oznacza perfekcyjne przewidywanie,
- **0** oznacza przewidywanie nie lepsze od losowego,
- **–1** oznacza całkowitą niezgodność między przewidywaniami a rzeczywistością.

MCC jest powszechnie uważany za bardziej informacyjną miarę niż dokładność, zwłaszcza w przypadku niezrównoważonych zbiorów danych. Zapewnia zrównoważoną ocenę jakości klasyfikacji obu klas, co czyni go cenną metryką przy ocenie modeli klasyfikacyjnych, w tym drzew decyzyjnych.

Decision Tree (Clasifier)

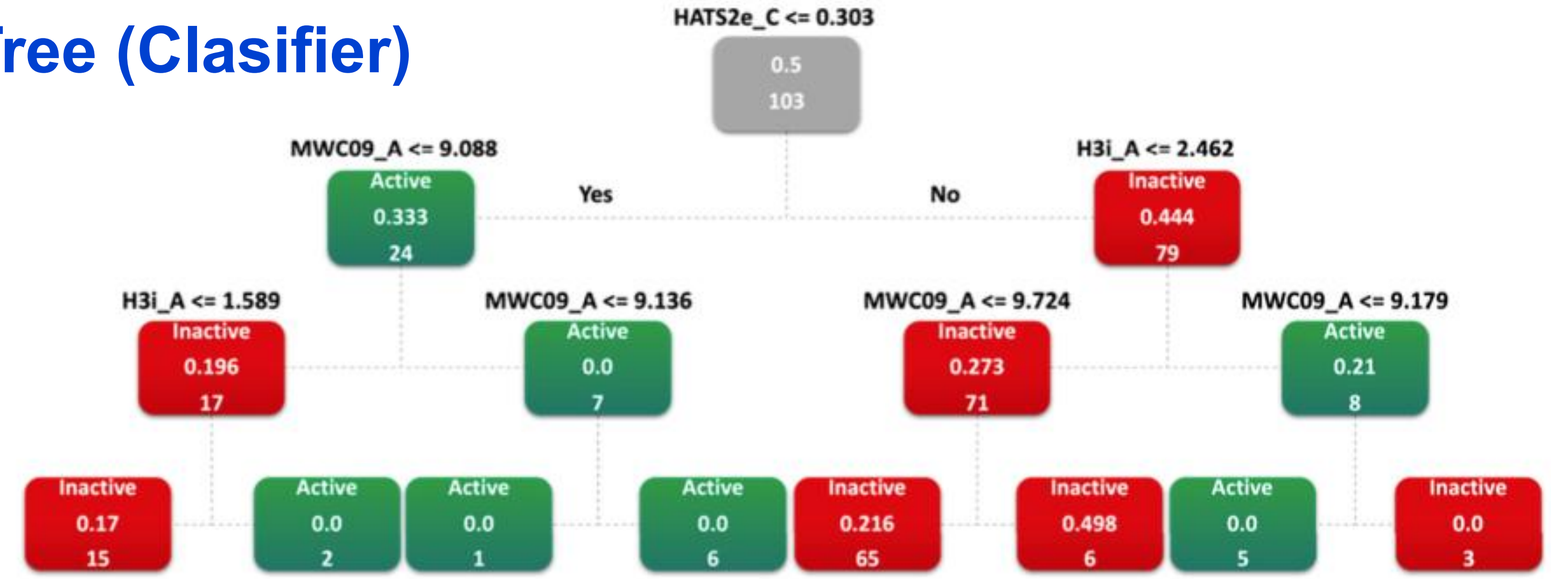


Figure 3. Classification Tree model for *Listeria monocytogenes* phage P100. (In nodes: the top row corresponds to the activity class predicted, the middle row corresponds to the gini index, and the bottom row corresponds to the number of samples assigned to the node. The green color of the node indicates the activity class "Active", while the red color indicates "Ina

Table 4. Summary of the Statistical Parameters Obtained from the P100, MS2, and Phi6 Models

Statistics	Model P100	Model MS2	Model Phi6
Accuracy for the training set	0.93	0.96	0.88
Accuracy for validation set	0.86	0.88	0.90
Cross-validation score (Train)	0.73	0.91	0.77
Cross-validation score (Test)	0.71	0.86	0.82

Table 5. Confusion Matrix for P100, MS2, and Phi6 Models[a]

		Predicted					
		P100		MS2		Phi6	
		Active	Inactive	Active	Inactive	Active	Inactive
Observed	Active	TP (38)	FN (4)	TP (40)	FN (3)	TP (28)	FN (2)
	Inactive	FP (3)	TN (6)	FP (3)	TN (5)	FP (5)	TN (18)

[a]The table includes the colors corresponding to the validity of the predictions made: green for valid, red for invalid. TP – True Positive; FN – False Negative; FP – False Positive; TN – True Negative.

Source: Zdybel S., et al. (2024) *J. Chem. Inf. Model. 2024, 64, 6, 1996–2007*

Decision Tree ”pruning”

•**Budowa „maksymalnego” drzewa** (optymalnie z pełną jednorodnością liści, tzn. w liściach znajdują się tylko obserwacje należące do jednej klasy).

•**Ocena poprawności klasyfikacji** w teście walidacji krzyżowej oraz w walidacji zewnętrznej.

•**Redukcja głębokości drzewa w procesie przycinania** (usuwanie końcowych gałęzi drzewa) oraz ponowna ocena poprawności klasyfikacji w teście walidacyjnym.

•**Dalsze przycinanie drzewa** (jeżeli jest to możliwe).

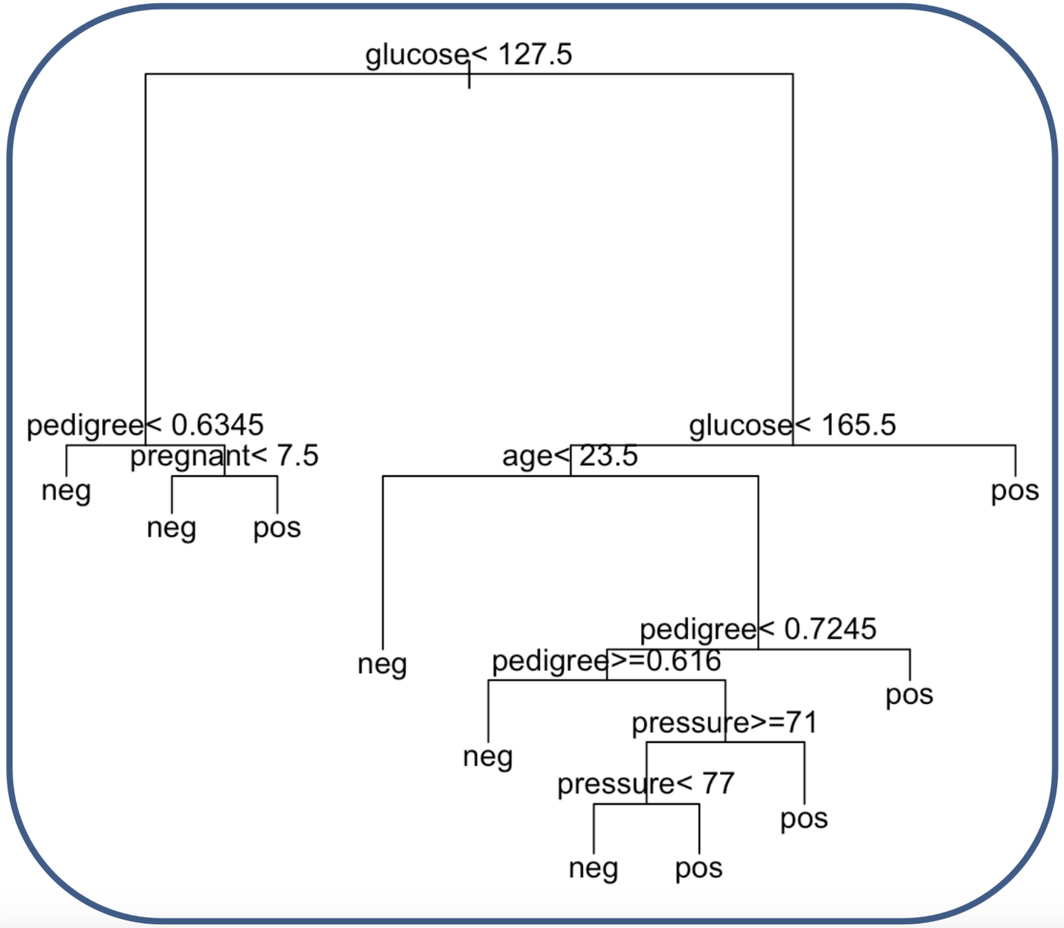
Przycinanie drzew (tree pruning) w drzewach decyzyjnych jest kluczową techniką zapobiegającą nadmiernemu dopasowaniu. Nadmierne dopasowanie występuje wtedy, gdy model zbyt dokładnie uczy się danych treningowych, wychwytując szum i nieistotne wzorce zamiast rzeczywistego sygnału. Prowadzi to do słabej zdolności uogólniania na nowe, niewidziane wcześniej dane.

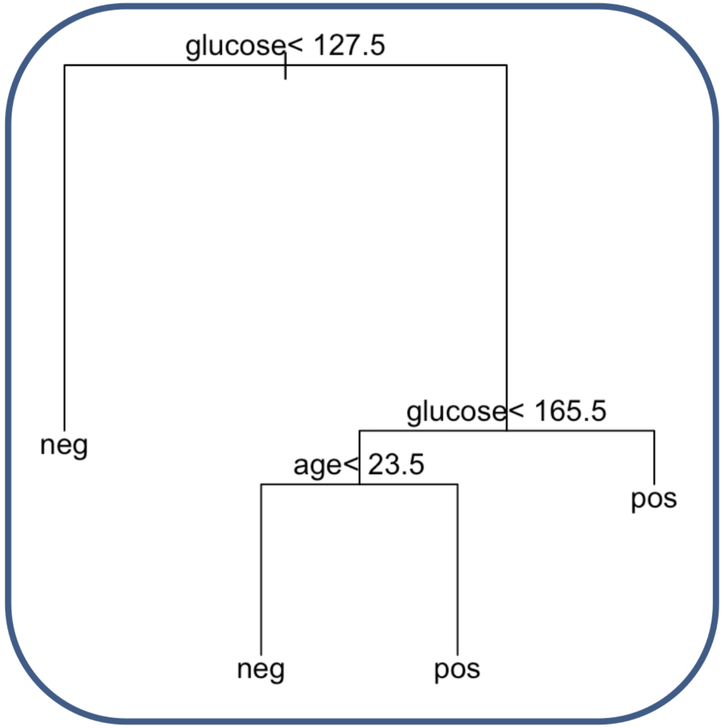
Celem przycinania jest uproszczenie drzewa decyzyjnego poprzez usunięcie tych jego części, które nie wnoszą istotnej wartości predykcyjnej. Można to osiągnąć na dwa główne sposoby:

Pre-pruning: polega na wcześniejszym zatrzymaniu procesu budowy drzewa, zanim stanie się ono zbyt złożone. Przykłady obejmują ustawienie maksymalnej głębokości drzewa lub minimalnej liczby próbek wymaganych do podziału węzła.
Post-pruning: polega na zbudowaniu pełnego drzewa, a następnie usunięciu gałęzi lub poddrzew, które nie poprawiają znacząco wyników na zbiorze walidacyjnym.
Dzięki redukcji złożoności drzewa decyzyjnego, przycinanie pomaga:

- **Poprawić uogólnianie:** model jest mniej podatny na nadmierne dopasowanie i lepiej radzi sobie z nowymi danymi.
- **Zwiększyć interpretowalność:** prostsze drzewa są łatwiejsze do zrozumienia i wizualizacji, co ułatwia interpretację procesu podejmowania decyzji przez model.
- **Podnieść efektywność:** mniejsze drzewa wymagają mniej zasobów obliczeniowych podczas trenowania i przewidywania.

Decision Tree ”pruning”





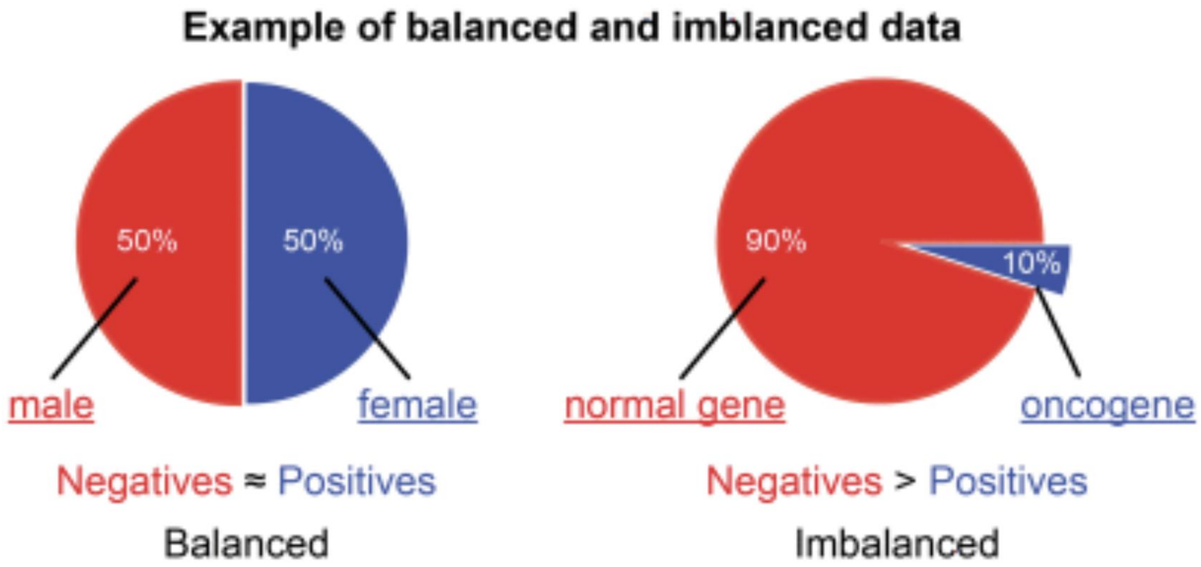
Przycinanie drzew decyzyjnych jest kluczową techniką zapobiegającą przeuczeniu. Przeuczenie występuje wtedy, gdy model zbyt dokładnie uczy się danych treningowych, wychwytując szum i nieistotne wzorce zamiast właściwego sygnału. Skutkuje to słabą zdolnością uogólniania na nowe, niewidziane wcześniej dane.

Celem przycinania jest uproszczenie drzewa decyzyjnego poprzez usunięcie tych jego części, które nie wnoszą istotnej wartości predykcyjnej. Można to osiągnąć na dwa główne sposoby:

- **Przycinanie wstępne (pre-pruning):** polega na wcześniejszym zatrzymaniu procesu budowy drzewa, zanim stanie się ono zbyt złożone. Przykładowo można ustawić ograniczenia na maksymalną głębokość drzewa lub minimalną liczbę próbek wymaganą do podziału węzła.
- **Przycinanie końcowe (post-pruning):** polega na zbudowaniu pełnego drzewa, a następnie usuwaniu gałęzi lub poddrzew, które nie poprawiają znacząco wyników na zbiorze walidacyjnym.

- Zmniejszając złożoność drzewa decyzyjnego, przycinanie pomaga: **Poprawić uogólnianie:** model jest mniej podatny na przeuczenie i lepiej radzi sobie z nowymi danymi.
- **Zwiększyć interpretowalność:** prostsze drzewa są łatwiejsze do zrozumienia i wizualizacji, co ułatwia interpretację procesu podejmowania decyzji przez model.
- **Zwiększyć efektywność:** mniejsze drzewa wymagają mniej zasobów obliczeniowych podczas trenowania i predykcji.

Unbalanced data



Niezrównoważone dane to takie zbiory danych, w których jedna klasa występuje znacznie częściej niż druga. Stanowi to wyzwanie dla wielu algorytmów klasyfikacji, w tym drzew decyzyjnych.

W przypadku pracy z niezrównoważonymi danymi standardowe metryki klasyfikacyjne, takie jak dokładność (accuracy), mogą być mylące. Model może bowiem osiągnąć wysoką dokładność, po prostu przewidując większościową klasę dla wszystkich obserwacji, bez faktycznego nauczenia się rzeczywistych wzorców.

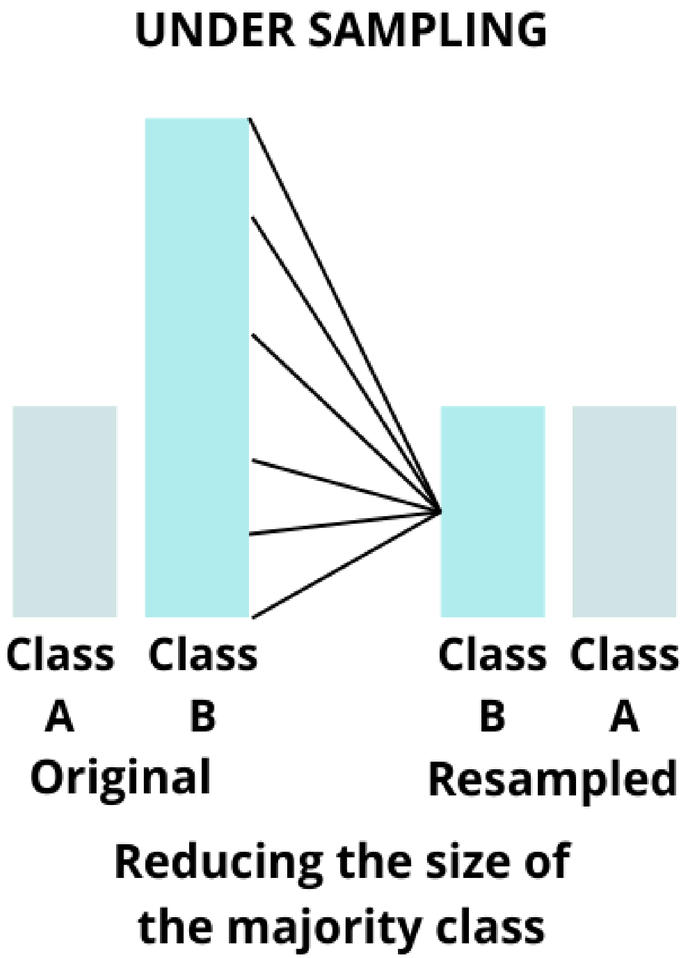
Aby poradzić sobie z tym problemem, można zastosować różne techniki:
Próbkowanie (resampling):

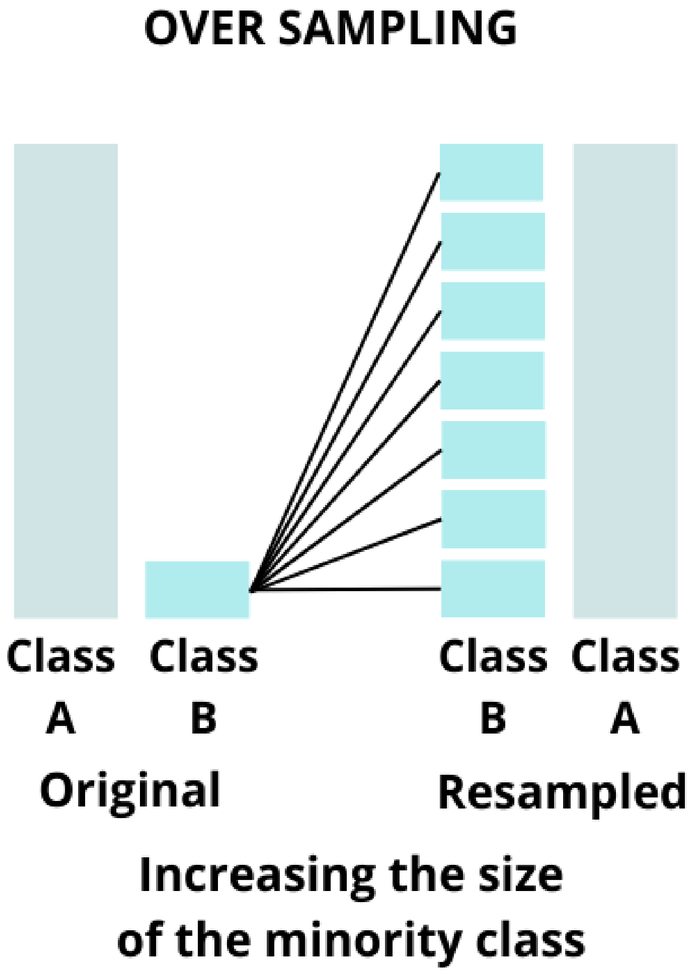
- **Oversampling:** zwiększenie liczby próbek klasy mniejszościowej poprzez duplikację istniejących przykładów lub generowanie nowych danych syntetycznych (np. przy użyciu SMOTE – Synthetic Minority Over-sampling Technique).
- **Undersampling:** zmniejszenie liczby próbek klasy większościowej, co jednak może prowadzić do utraty cennych informacji.

Uczenie kosztoczułe (cost-sensitive learning): nadawanie różnych kosztów błędnym klasyfikacjom poszczególnych klas. Na przykład wyższa kara za błędną klasyfikację próbki z klasy mniejszościowej.
Używanie odpowiednich metryk oceny: skupienie się na metrykach mniej podatnych na niezrównoważenie klas, takich jak precision, recall, F1-score czy AUC-ROC.

Dzięki odpowiedniemu podejściu do problemu niezrównoważenia klas można poprawić wydajność modeli klasyfikacyjnych pracujących na rzeczywistych zbiorach danych i uzyskać bardziej wiarygodne predykcje.

Unbalanced data





Unbalanced data problem

Losowe generowanie podobnych danych dla klasy mniejszościowej:

- Metoda **SMOTE** (Synthetic Minority Oversampling Technique) generuje nowe próbki pomiędzy istniejącymi punktami danych, bazując na ich lokalnej gęstości.
- **Zmniejszanie wag (down-weighting):** redukowanie wagi obiektów należących do klasy większościowej (analogiczne do undersamplingu).
- **Zwiększanie wag (up-weighting):** zwiększanie wagi obiektów należących do klasy mniejszościowej (analogiczne do oversamplingu).

Sprawozdanie

Regresja (DT):

1.Wykonać model **drzewa decyzyjnego** z całą pulą deskryptorów.
2.Wykonać **feature selection** z podstawowymi parametrami dla algorytmu Backward.
3.Przygotować **optymalizację maksymalnej głębokości drzewa** z wykresem statystyk generowanych przez model dla różnych głębokości.
4.Przygotować **optymalizację minimalnej liczby próbek w liściu** (min_samples_leaf) z wykresem statystyk generowanych przez model dla różnych wartości.
5.Wykonać jednoczesną optymalizację głębokości drzewa i minimalnej liczby próbek w liściu z odpowiadającymi im statystykami.
6.Powtórzyć pkt 1-5 dla **drzewa decyzyjnego z przycinaniem (pruning)** lub ustawieniem minimalnego kryterium czystości liści.
7.Napisać wnioski wynikające z powyższych analiz z uwzględnieniem tego, które deskryptory są najczęściej wybierane do różnych modeli.

Sprawozdanie

Klasyfikacja (DT):

1.Wykonać model **drzewa decyzyjnego** z całą pulą deskryptorów.
2.Wykonać **feature selection** z podstawowymi parametrami dla algorytmu Backward.
3.Przygotować **optymalizację maksymalnej głębokości drzewa** z wykresem statystyk generowanych przez model dla różnych głębokości.
4.Przygotować **optymalizację minimalnej liczby próbek w liściu** (min_samples_leaf) z wykresem statystyk generowanych przez model dla różnych wartości.
5.Wykonać jednoczesną optymalizację głębokości drzewa i minimalnej liczby próbek w liściu z odpowiadającymi im statystykami.
6.Powtórzyć pkt 1-5 dla **drzewa decyzyjnego z przycinaniem (pruning)** lub ustawieniem minimalnego kryterium czystości liści.
7.Napisać wnioski wynikające z powyższych analiz z uwzględnieniem tego, które deskryptory są najczęściej wybierane do różnych modeli.