POLITECHNIKA RZESZOWSKA

im. Ignacego Łukasiewicza

WYDZIAŁ MATEMATYKI I FIZYKI STOSOWANEJ

Wnioskowanie w warunkach niepewności - sieci Bayesa

Analiza sieci Bayesowskich na zbiorze danych o winie

Stachiewicz Dawid

Inżynieria i Analiza Danych, III rok
Grupa laboratoryjna nr 5, Nr albumu: 173218

Rzeszów, 31 maja 2025

Spis treści

1 Wprowadzenie

Celem niniejszego projektu jest zastosowanie sieci Bayesowskich do modelowania zależności między właściwościami chemicznymi wina a jego jakością. Analiza została przeprowadzona na podstawie ogólnodostępnego zbioru danych *Wine Quality Dataset*, który obejmuje chemiczne i sensoryczne właściwości czerwonych win portugalskich typu *Vinho Verde*.

Zbiór danych zawiera 1599 obserwacji oraz 12 zmiennych — 11 wejściowych (opisujących cechy fizykochemiczne) i 1 wyjściową, będącą oceną jakości wina na skali od 0 do 10. Dane te mogą być traktowane zarówno jako problem klasyfikacyjny, jak i regresyjny, przy czym klasy nie są równomiernie reprezentowane (znaczna przewaga win przeciętnych względem wybitnych lub słabych).

Analiza obejmuje następujące etapy:

- przygotowanie i dyskretyzację danych wejściowych (w tym transformację zmiennych ciągłych na kategorie),
- przeprowadzenie testów istotności statystycznej (test χ^2) w celu wstępnej analizy zależności,
- budowę i porównanie różnych struktur sieci Bayesowskich (w tym HC, GS, PC, IAMB i innych),
- ocenę jakości modeli za pomocą kryterium informacyjnego BIC,
- estymację parametrów warunkowych dla wybranej najlepszej sieci,
- oraz symulację scenariuszy predykcyjnych z wykorzystaniem dopasowanego modelu probabilistycznego.

Zbiór danych pochodzi ze strony Kaggle. Dane zostały pierwotnie zebrane i opracowane przez P. Cortez i wsp., a ich szczegółowy opis został opublikowany w artykule *Modeling wine preferences by data mining from physicochemical properties* (2009).

Głównym celem projektu jest stworzenie przejrzystego i poprawnego modelu probabilistycznego, który umożliwia zarówno interpretację zależności między cechami chemicznymi a jakością wina, jak i predykcję jakości na podstawie zaobserwowanych właściwości.

2 Opis danych

Dane analizowane w projekcie pochodzą ze zbioru `Wine Quality Dataset` i dotyczą czerwonych win portugalskich. Zbiór składa się z **1599 obserwacji**, z których każda opisuje jedną próbkę wina za pomocą zestawu **11 zmiennych wejściowych** (fizykochemicznych) oraz jednej zmiennej wyjściowej — `quality` — oznaczającej jakość wina w skali od 0 do 10 (na potrzeby analizy została ona później zdyskretyzowana do trzech poziomów: *Niska*, *Średnia*, *Wysoka*).

Zmienne wejściowe można podzielić na kilka kategorii:

- **Parametry kwasowości:** `fixed.acidity`, `volatile.acidity`, `citric.acid`,
- **Zawartość cukru i alkoholu:** `residual.sugar`, `alcohol`,
- **Zawartość soli i siarki:** `chlorides`, `free.sulfur.dioxide`, `total.sulfur.dioxide`,
- **Parametry ogólne:** `density`, `pH`, `sulphates`.

Poniżej przedstawiono przykładowy fragment struktury danych zaraz po ich wczytaniu w środowisku `R`. Jak widać, zmienne są początkowo w postaci ciągłej (numerycznej), co wymagało późniejszej dyskretyzacji, niezbędnej do konstrukcji sieci Bayesowskich:

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
1	7.4	0.700	0.00	1.90	0.076	11	34	0.99780	3.51	0.56	9.4	5
2	7.8	0.880	0.00	2.60	0.098	25	67	0.99680	3.20	0.68	9.8	5
3	7.8	0.760	0.04	2.30	0.092	15	54	0.99700	3.26	0.65	9.8	5
4	11.2	0.280	0.56	1.90	0.075	17	60	0.99800	3.16	0.58	9.8	6
5	7.4	0.700	0.00	1.90	0.076	11	34	0.99780	3.51	0.56	9.4	5
6	7.4	0.660	0.00	1.80	0.075	13	40	0.99780	3.51	0.56	9.4	5
7	7.9	0.600	0.06	1.60	0.069	15	59	0.99640	3.30	0.46	9.4	5
8	7.3	0.650	0.00	1.20	0.065	15	21	0.99460	3.39	0.47	10.0	7
9	7.8	0.580	0.02	2.00	0.073	9	18	0.99680	3.36	0.57	9.5	7
10	7.5	0.500	0.36	6.10	0.071	17	102	0.99780	3.35	0.80	10.5	5
11	6.7	0.580	0.08	1.80	0.097	15	65	0.99590	3.28	0.54	9.2	5
12	7.5	0.500	0.36	6.10	0.071	17	102	0.99780	3.35	0.80	10.5	5
13	5.6	0.615	0.00	1.60	0.089	16	59	0.99430	3.58	0.52	9.9	5
14	7.8	0.610	0.29	1.60	0.114	9	29	0.99740	3.26	1.56	9.1	5
15	8.9	0.620	0.18	3.80	0.176	52	145	0.99860	3.16	0.88	9.2	5
16	8.9	0.620	0.19	3.90	0.170	51	148	0.99860	3.17	0.93	9.2	5
17	8.5	0.280	0.56	1.80	0.092	35	103	0.99690	3.30	0.75	10.5	7
18	8.1	0.560	0.28	1.70	0.368	16	56	0.99680	3.11	1.28	9.3	5
19	7.4	0.590	0.08	4.40	0.086	6	29	0.99740	3.38	0.50	9.0	4
20	7.9	0.320	0.51	1.80	0.341	17	56	0.99690	3.04	1.08	9.2	6
21	8.9	0.220	0.48	1.80	0.077	29	60	0.99680	3.39	0.53	9.4	6
22	7.6	0.390	0.31	2.30	0.082	23	71	0.99820	3.52	0.65	9.7	5
23	7.9	0.430	0.21	1.60	0.106	10	37	0.99660	3.17	0.91	9.5	5
24	8.5	0.490	0.11	2.30	0.084	9	67	0.99680	3.17	0.53	9.4	5
25	6.9	0.400	0.14	2.40	0.085	21	40	0.99680	3.43	0.63	9.7	6
26	6.3	0.390	0.16	1.40	0.080	11	23	0.99550	3.34	0.56	9.3	5

Rysunek 1: Fragment danych po wczytaniu w R

Każdy wiersz w zbiorze danych odpowiada jednej próbce wina, a każda kolumna reprezentuje określoną cechę fizykochemiczną lub ocenę jakości. Dane są gotowe do dalszej eksploracji i przetwarzania, ale wymagają transformacji cech liczbowych na zmienne kategoryczne (np. poziomy: niski, średni, wysoki) przed ich wykorzystaniem w modelach probabilistycznych.

3 Obróbka danych: dyskretyzacja

Dyskretyzacja jest jednym z kluczowych etapów przygotowania danych do analizy z użyciem sieci Bayesowskich. Ponieważ modele te operują na zmiennych kategorialnych, konieczne było przekształcenie wszystkich zmiennych ciągłych (numerycznych) na zmienne dyskretne z ustalonymi progami podziału.

Dla każdej cechy ustalono trójstopniową skalę: *Niska*, *Średnia*, *Wysoka*. Przedziały zostały dobrane na podstawie analizy rozkładu wartości zmiennych, w oparciu o wiedzę ekspercką oraz konsultacje z istniejącymi projektami wykorzystującymi ten sam zbiór danych.

Przykładowe progi zastosowane podczas dyskretyzacji:

- `alcohol`: **Niska** < 10, **Średnia** $= 10$–12, **Wysoka** > 12
- `fixed.acidity`: **Niska** < 6.5, **Średnia** $= 6.5$–8.5, **Wysoka** > 8.5
- `volatile.acidity`: **Niska** < 0.3, **Średnia** $= 0.3$–0.6, **Wysoka** > 0.6
- `citric.acid`: **Niska** < 0.2, **Średnia** $= 0.2$–0.4, **Wysoka** > 0.4
- `residual.sugar`: **Niski** < 2, **Średni** $= 2$–5, **Wysoki** > 5
- `chlorides`: **Niska** < 0.07, **Średnia** $= 0.07$–0.1, **Wysoka** > 0.1
- `free.sulfur.dioxide`: **Niski** < 15, **Średni** $= 15$–30, **Wysoki** > 30
- `total.sulfur.dioxide`: **Niski** < 40, **Średni** $= 40$–100, **Wysoki** > 100
- `density`: **Niska** < 0.995, **Średnia** $= 0.995$–0.998, **Wysoka** > 0.998
- `pH`: **Niskie** < 3.2, **Średnie** $= 3.2$–3.4, **Wysokie** > 3.4
- `sulphates`: **Niska** < 0.5, **Średnia** $= 0.5$–0.7, **Wysoka** > 0.7
- `quality`: **Niska** $= 0$–5, **Średnia** $= 5$–7, **Wysoka** $= 7$–10

Dyskretyzacja pozwala uprościć analizę oraz umożliwia bezpośrednie zastosowanie narzędzi w pakiecie `bnlearn`, które wymagają danych w postaci kategorii. Poniżej przedstawiono przykładowy efekt dyskretyzacji zmiennych:

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
1	Średnia	Wysoka	Niska	Niski	Średnia	Niski	Niski	Średnia	Wysokie	Średnia	Niska	Niska
2	Średnia	Wysoka	Niska	Średni	Średnia	Średni	Średni	Średnia	Niskie	Średnia	Niska	Niska
3	Średnia	Wysoka	Niska	Średni	Średnia	Niski	Średni	Średnia	Średnie	Średnia	Niska	Niska
4	Wysoka	Niska	Wysoka	Niski	Średnia	Średni	Średni	Średnia	Niskie	Średnia	Niska	Średnia
5	Średnia	Wysoka	Niska	Niski	Średnia	Niski	Niski	Średnia	Wysokie	Średnia	Niska	Niska
6	Średnia	Wysoka	Niska	Niski	Średnia	Niski	Niski	Średnia	Wysokie	Średnia	Niska	Niska
7	Średnia	Średnia	Niska	Niski	Niska	Niski	Średni	Średnia	Średnie	Niska	Niska	Niska
8	Średnia	Wysoka	Niska	Niski	Niska	Niski	Niski	Niska	Średnie	Niska	Niska	Średnia
9	Średnia	Średnia	Niska	Niski	Średnia	Niski	Niski	Średnia	Średnie	Średnia	Niska	Średnia
10	Średnia	Średnia	Średnia	Wysoki	Średnia	Średni	Wysoki	Średnia	Średnie	Wysoka	Średnia	Niska
11	Średnia	Średnia	Niska	Niski	Średnia	Niski	Średni	Średnia	Średnie	Średnia	Niska	Niska
12	Średnia	Średnia	Średnia	Wysoki	Średnia	Średni	Wysoki	Średnia	Średnie	Wysoka	Średnia	Niska
13	Niska	Wysoka	Niska	Niski	Średnia	Średni	Średni	Niska	Wysokie	Średnia	Niska	Niska
14	Średnia	Wysoka	Średnia	Niski	Wysoka	Niski	Niski	Średnia	Średnie	Wysoka	Niska	Niska
15	Wysoka	Wysoka	Niska	Średni	Wysoka	Wysoki	Wysoki	Wysoka	Niskie	Wysoka	Niska	Niska
16	Wysoka	Wysoka	Niska	Średni	Wysoka	Wysoki	Wysoki	Wysoka	Niskie	Wysoka	Niska	Niska
17	Średnia	Niska	Wysoka	Niski	Średnia	Wysoki	Wysoki	Średnia	Średnie	Wysoka	Średnia	Średnia
18	Średnia	Średnia	Średnia	Niski	Wysoka	Średni	Średni	Średnia	Niskie	Wysoka	Niska	Niska
19	Średnia	Średnia	Niska	Średni	Średnia	Niski	Niski	Średnia	Średnie	Niska	Niska	Niska
20	Średnia	Średnia	Wysoka	Niski	Wysoka	Średni	Średni	Średnia	Niskie	Wysoka	Niska	Średnia
21	Wysoka	Niska	Wysoka	Niski	Średnia	Średni	Średni	Średnia	Średnie	Średnia	Niska	Średnia
22	Średnia	Średnia	Średnia	Średni	Średnia	Średni	Średni	Wysoka	Wysokie	Średnia	Niska	Niska
23	Średnia	Średnia	Średnia	Niski	Wysoka	Niski	Niski	Średnia	Niskie	Wysoka	Niska	Niska
24	Średnia	Średnia	Niska	Średni	Średnia	Niski	Średni	Średnia	Niskie	Średnia	Niska	Niska
25	Średnia	Średnia	Niska	Średni	Średnia	Średni	Niski	Średnia	Wysokie	Średnia	Niska	Średnia
26	Niska	Średnia	Niska	Niski	Średnia	Niski	Niski	Średnia	Średnie	Średnia	Niska	Niska

Rysunek 2: Efekty dyskretyzacji zmiennych — dane po przekształceniu do postaci kategorialnej

Po dokonaniu dyskretyzacji, wszystkie zmienne zostały również jawnie przekonwertowane do typu `factor`, aby zapewnić pełną kompatybilność z funkcjami pakietu `bnlearn`.

4 Analiza istotności: test chi-kwadrat

Aby zbadać zależności pomiędzy poszczególnymi zmiennymi w zbiorze danych, przeprowadzono test niezależności chi-kwadrat (χ^2) dla każdej pary zmiennych. Celem było wykrycie statystycznie istotnych relacji, które mogą sugerować potencjalne powiązania przyczynowo-skutkowe, a także dostarczyć informacji pomocnych przy konstruowaniu sieci Bayesowskich.

Test chi-kwadrat pozwala sprawdzić, czy dwie zmienne kategorialne są od siebie statystycznie niezależne. Hipotezy testu sformułowano następująco:

- H_0: Zmienne są niezależne (brak istotnego związku).
- H_1: Zmienne są zależne (istnieje istotna zależność).

Dla każdej pary zmiennych obliczono wartość p-value. Jeśli p-value było mniejsze niż ustalony poziom istotności $\alpha = 0{,}05$, to hipotezę zerową H_0 odrzucano na korzyść hipotezy alternatywnej H_1, co oznacza istnienie statystycznie istotnej zależności.

Wszystkie obliczenia wykonano automatycznie w pętli `for`, generując tabelę wyników z nazwami zmiennych, wartością p-value i informacją, czy zależność jest istotna. Poniżej zaprezentowano wizualizację najistotniejszych zależności według testu χ^2:

	Zmienna_1	Zmienna_2	p_value	Istotna
46	free.sulfur.dioxide	total.sulfur.dioxide	1.941991e-149	TRUE
7	fixed.acidity	density	2.140988e-142	TRUE
8	fixed.acidity	pH	9.144613e-131	TRUE
2	fixed.acidity	citric.acid	7.111000e-117	TRUE
59	density	alcohol	4.068040e-84	TRUE
27	citric.acid	pH	2.456893e-78	TRUE
12	volatile.acidity	citric.acid	2.620479e-78	TRUE
66	alcohol	quality	4.365631e-67	TRUE
41	chlorides	density	3.586244e-59	TRUE
34	residual.sugar	density	1.080862e-39	TRUE
26	citric.acid	density	2.016848e-37	TRUE
28	citric.acid	sulphates	3.149382e-31	TRUE
44	chlorides	alcohol	8.000292e-31	TRUE
57	density	pH	1.011330e-30	TRUE
65	sulphates	quality	1.659464e-30	TRUE
21	volatile.acidity	quality	6.388692e-30	TRUE
10	fixed.acidity	alcohol	9.614153e-24	TRUE
4	fixed.acidity	chlorides	1.187514e-21	TRUE
56	total.sulfur.dioxide	quality	6.005617e-19	TRUE
19	volatile.acidity	sulphates	3.060413e-18	TRUE
1	fixed.acidity	volatile.acidity	3.597093e-16	TRUE
60	density	quality	1.061018e-15	TRUE
20	volatile.acidity	alcohol	3.197530e-15	TRUE
64	sulphates	alcohol	1.007365e-14	TRUE
25	citric.acid	total.sulfur.dioxide	8.907493e-14	TRUE
42	chlorides	pH	1.953245e-13	TRUE

Rysunek 3: Najistotniejsze zależności między zmiennymi wg testu χ^2 (najniższe wartości p-value)

Test chi-kwadrat był również pomocny w identyfikacji zmiennych silnie skorelowanych z jakością wina (`quality`), takich jak `alcohol`, `sulphates` oraz `volatile.acidity`, które później okazały się kluczowymi elementami w konstrukcji sieci Bayesowskiej.

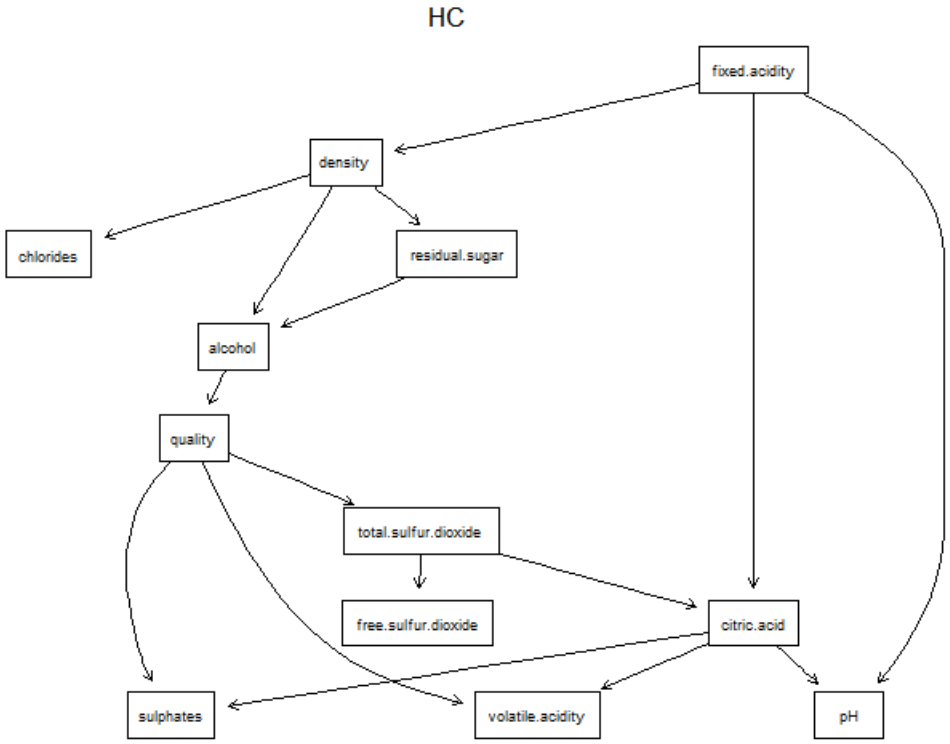
5 Tworzenie sieci Bayesowskich

W ramach projektu skonstruowano wiele sieci Bayesowskich, wykorzystując zarówno algorytmy optymalizacji struktury (takie jak Hill Climbing oraz Tabu Search), jak i podejścia oparte na testach niezależności (PC-Stable, Grow-Shrink, IAMB i jego warianty).

Dla algorytmów opartych na testach niezależności (PC-Stable, Grow-Shrink, IAMB, Fast IAMB, Inter IAMB, IAMB-FDR) konieczne było przeanalizowanie spójności struktur grafowych, identyfikacja niesymetrycznych łuków oraz w niektórych przypadkach — ręczna korekta sieci w celu zapewnienia poprawności (tj. struktury typu DAG — Directed Acyclic Graph).

Poniżej przedstawiono kluczowe przykłady sieci oraz ich wersje po wprowadzeniu poprawek:

5.1 Sieć Hill Climbing (HC)



Rysunek 4: Sieć Bayesowska skonstruowana algorytmem Hill Climbing (HC)

Sieć Bayesowska wygenerowana metodą Hill Climbing (HC) stanowi wynik optymalizacji funkcji kryterialnej BIC, której celem jest znalezienie najlepiej dopasowanej struktury grafowej dla zbioru danych.

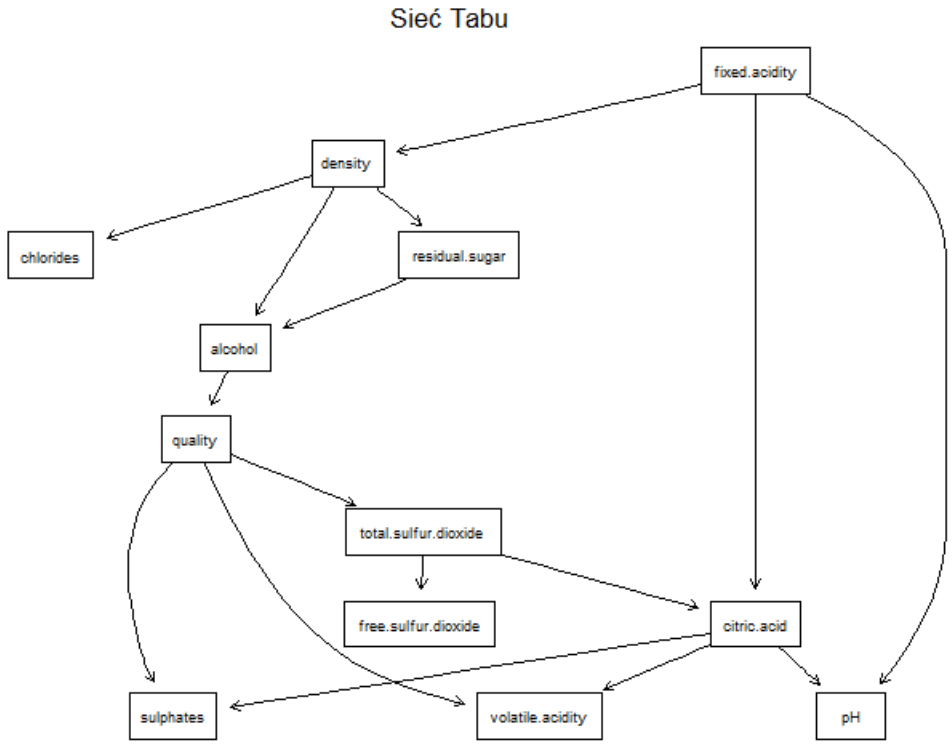
W analizowanej strukturze można zauważyć logiczne powiązania pomiędzy zmiennymi chemicznymi:

- `alcohol`, `volatile.acidity` oraz `sulphates` są bezpośrednio połączone z `quality`, co sugeruje ich kluczowy wpływ na ocenę jakości wina.
- `density` wpływa na `alcohol`, co jest zgodne z faktem, że zawartość alkoholu ma wpływ na gęstość cieczy.

- Powiązania między `fixed.acidity`, `citric.acid` i `pH` odzwierciedlają wewnętrzne relacje pomiędzy właściwościami kwasowymi napoju.

Struktura ta została uznana za najbardziej trafną pod względem informacyjnym (najniższa wartość BIC spośród porównywanych modeli), a jednocześnie jest ona zgodna z wiedzą dziedzinową dotyczącą chemii wina. Sieć ta została zatem wybrana jako podstawa do dalszej estymacji parametrów i budowy scenariuszy predykcyjnych.

5.2 Sieć Tabu



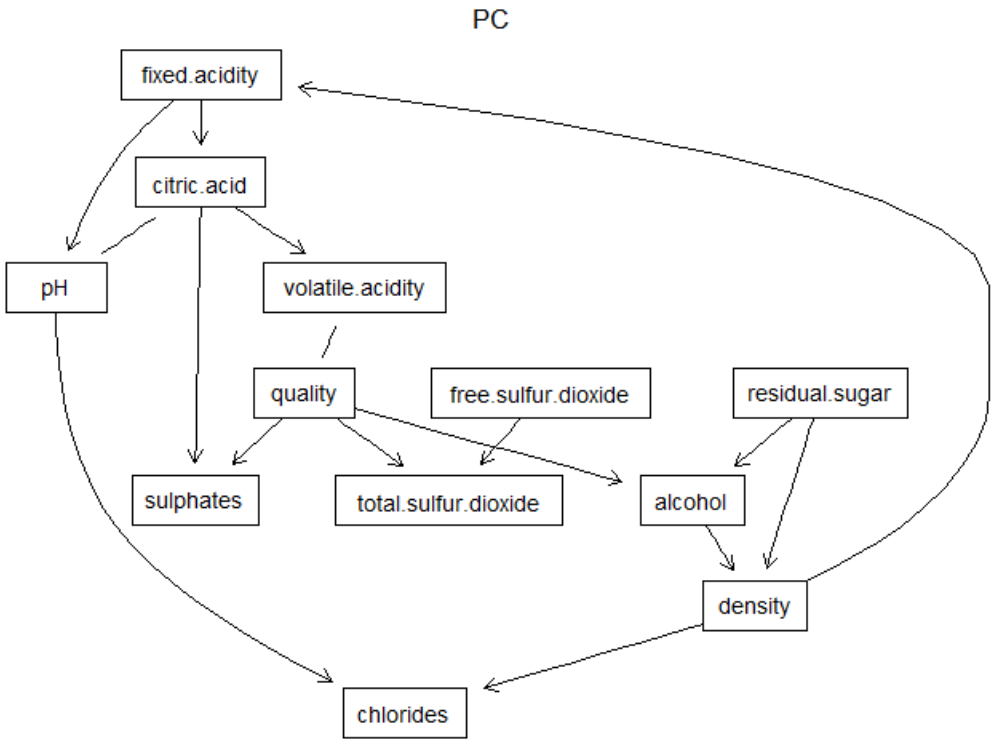
Rysunek 5: Sieć Bayesowska skonstruowana algorytmem Tabu Search

Algorytm Tabu Search, podobnie jak Hill Climbing, przeszukuje przestrzeń możliwych sieci, ale w odróżnieniu od klasycznego HC, przechowuje listę zakazanych (tabu) operacji, aby uniknąć zapętlenia się w lokalnych minimach. W rezultacie umożliwia dotarcie do lepszych rozwiązań globalnych.

W zaprezentowanej sieci możemy zaobserwować zależności zbliżone do HC, m.in. wpływ zmiennych `alcohol` i `sulphates` na `quality`. Jednak w porównaniu z HC sieć ta zawiera dodatkowe połączenia między zmiennymi pośrednimi — np. między `pH` a `citric.acid`, czy między `chlorides` a `density`, co może świadczyć o bardziej złożonym podejściu do odwzorowania fizykochemicznych związków.

Mimo że wynik BIC dla tego modelu był zbliżony do HC, to jego struktura okazała się nieco mniej czytelna i wymagałaby dalszych analiz w celu eliminacji potencjalnych połączeń bez uzasadnienia chemicznego.

5.3 Sieć PC-Stable (przed poprawką)

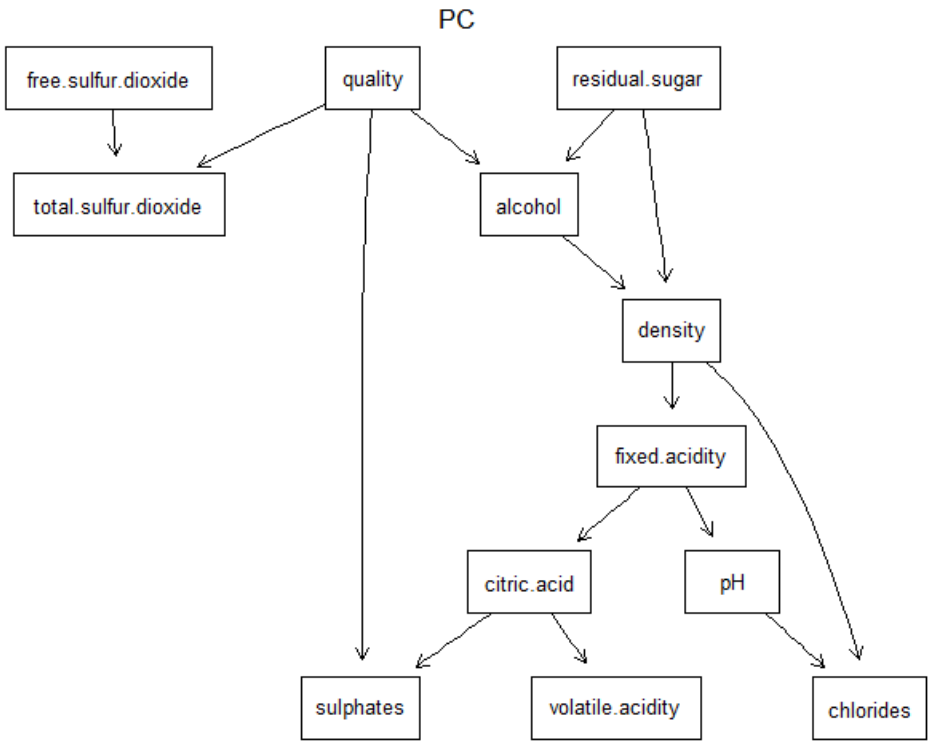


Rysunek 6: Sieć PC-Stable przed poprawkami (niesymetryczne łuki, brak spójności)

Algorytm PC-Stable bazuje na analizie testów niezależności statystycznej i generuje graf szkieletowy, który następnie orientuje. W wersji przed poprawką sieć ta posiadała niesymetryczne łuki oraz niespójności, które uniemożliwiały bezpośrednie wykorzystanie jej do estymacji prawdopodobieństw (sieć nie była DAGiem).

Występowały również izolowane wierzchołki i błędnie skierowane zależności, np. od `quality` do `volatile.acidity`, co nie znajduje uzasadnienia ani w literaturze, ani w fizyce chemicznej.

5.4 Sieć PC-Stable (po poprawkach)

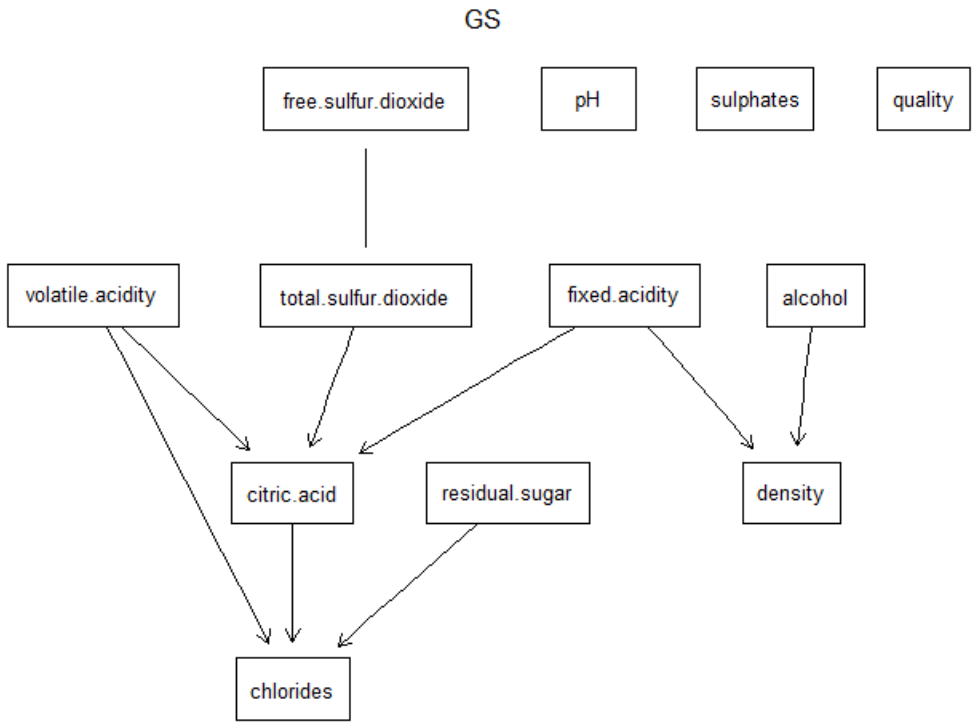


Rysunek 7: Sieć PC-Stable po ręcznych poprawkach – struktura DAG

Po ręcznych modyfikacjach, polegających na usunięciu niezgodnych łuków i uzupełnieniu brakujących połączeń, sieć PC-Stable zyskała strukturę zgodną z wymogami grafu acyklicznego (DAG).

Zmieniono kierunki łuków zgodnie z logicznym i przyczynowym porządkiem zmiennych (np. `alcohol` wpływa na `quality`, a nie odwrotnie). Poprawiona wersja nadaje się do dalszej analizy, chociaż jej wynik BIC okazał się gorszy w porównaniu z siecią HC.

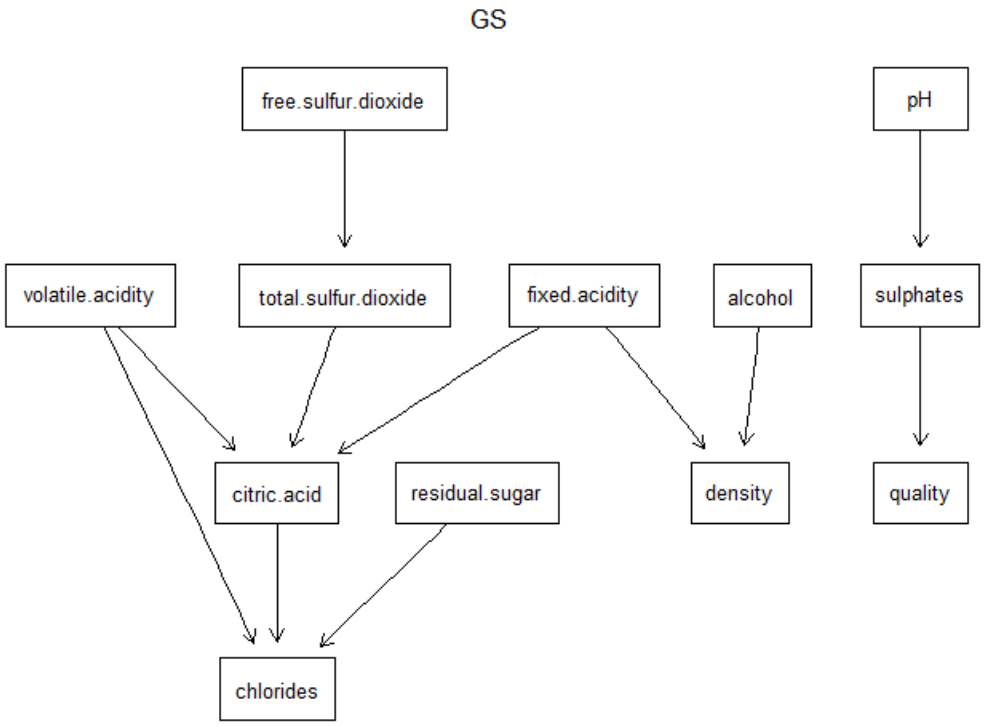
5.5 Sieć Grow-Shrink (GS) – przed poprawką



Rysunek 8: Sieć Grow-Shrink (GS) przed korektą struktury

Algorytm Grow-Shrink jest jednym z najprostszych algorytmów opartych na testach niezależności. W wersji pierwotnej wygenerował sieć zawierającą niesymetryczne łuki oraz niepełne połączenia między kluczowymi zmiennymi. Występowały izolowane węzły oraz brakowało połączeń, które pojawiały się w innych modelach, np. między `sulphates` a `quality`.

5.6 Sieć Grow-Shrink (GS) – po poprawce



Rysunek 9: Sieć Grow-Shrink (GS) po poprawkach — spójna struktura bez łuków zwrotnych

Po zastosowaniu ręcznych poprawek, sieć GS zyskała spójną strukturę DAG. Uzupełniono połączenia między ważnymi zmiennymi, np. dodano łuk z `sulphates` do `quality`, a także poprawiono kierunki łuków tam, gdzie występowały niesymetryczne relacje.

Choć wynik BIC tego modelu nie był najlepszy, jego interpretowalność i zgodność z rzeczywistością chemiczną danych znacznie się poprawiła. Model ten mógłby posłużyć jako alternatywa do dalszej analizy, gdyby nieco lepiej spełniał kryteria statystyczne.

Pozostałe sieci

Oprócz powyższych, w projekcie zbudowano także sieci przy pomocy algorytmów:

- IAMB (Incremental Association Markov Blanket),
- Fast IAMB,
- Inter IAMB (Interleaved IAMB),
- IAMB-FDR (z kontrolą fałszywego odkrycia).

Dla każdej z tych sieci przeprowadzono ręczną analizę poprawności struktury (w tym usunięcie niesymetrycznych łuków oraz zapewnienie pełnej spójności grafu). Dodatkowo każda z tych sieci została oceniona przy użyciu kryterium BIC, a także zwizualizowana przed i po korekcie.

Z powodu ograniczonej objętości raportu, nie zamieszczono osobnych podrozdziałów i wykresów dla tych modeli, jednak ich konstrukcja, obróbka i analiza przebiegały analogicznie jak w przypadku sieci PC i GS.

6 Porównanie modeli BIC

Dla każdej ze skonstruowanych sieci obliczono wartość kryterium informacyjnego BIC (Bayesian Information Criterion), które pozwala ocenić jakość dopasowania modelu do danych przy uwzględnieniu liczby parametrów. Niższa wartość BIC oznacza lepszy kompromis między złożonością modelu a jego trafnością predykcyjną.

Wyniki porównania BIC przedstawiono na wykresie:

	Model	BIC
6	Hill Climbing (HC)	-15626.80
7	Tabu	-15626.80
8	PC	-15719.88
2	IAMB	-16574.38
3	IAMB FDR	-16588.72
1	Grow-Shrink (GS)	-16611.73
4	Inter IAMB	-16664.76
5	Fast IAMB	-17250.84
9	Ręczna	-17503.13

Rysunek 10: Porównanie wartości BIC dla różnych algorytmów konstrukcji sieci

Z analizy wynika, że sieć uzyskana przy użyciu algorytmu Hill Climbing osiągnęła najniższą wartość BIC, co czyni ją najlepszym kandydatem do dalszej analizy, estymacji parametrów i budowy scenariuszy predykcyjnych.

7 Wybór najlepszego modelu

W celu porównania jakości różnych modeli sieci Bayesowskich zastosowano kryterium informacyjne BIC (Bayesian Information Criterion). Kryterium to uwzględnia zarówno stopień dopasowania modelu do danych, jak i jego złożoność, penalizując sieci o dużej liczbie parametrów.

Spośród wszystkich przetestowanych modeli (HC, Tabu, PC, GS, IAMB, Fast IAMB, Inter IAMB, IAMB-FDR, model ręczny), najniższą wartość BIC — a tym samym najlepszy kompromis między dokładnością a prostotą — osiągnął model zbudowany za pomocą algorytmu **Hill Climbing (HC)**.

- Algorytm HC tworzy strukturę sieci w sposób iteracyjny, wybierając lokalnie najlepsze modyfikacje (dodanie, usunięcie lub odwrócenie łuku), aż do osiągnięcia optimum globalnego względem przyjętego kryterium (tu: BIC).
- Ostateczny wynik BIC dla tego modelu był lepszy niż dla modeli zbudowanych zarówno metodami opartymi na testach niezależności (takimi jak PC-Stable czy GS), jak i metodami heurystycznymi (Tabu).

Z tego powodu model HC został wybrany jako **model referencyjny** do dalszej analizy. W kolejnych etapach raportu na jego podstawie:

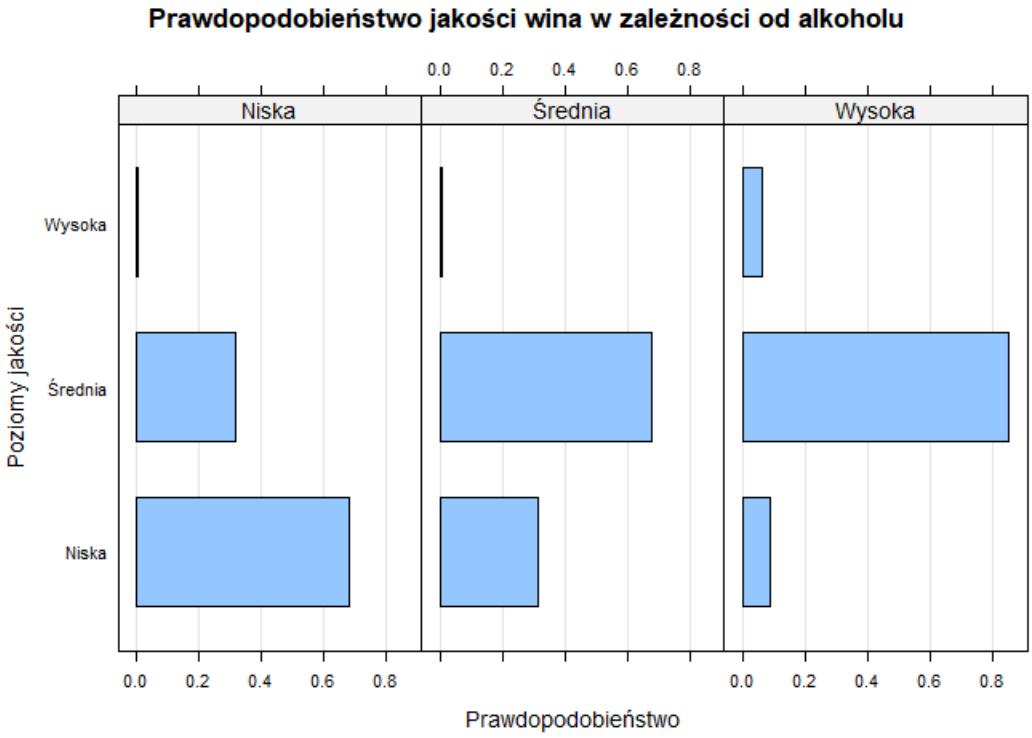
1. przeprowadzono **estymację parametrów warunkowych** (CPT),
2. wykonano **wizualizację rozkładów prawdopodobieństwa**,
3. stworzono i przeanalizowano **scenariusze predykcyjne**.

Wybór ten znajduje również uzasadnienie praktyczne — sieć HC posiadała przejrzystą strukturę, pozbawioną błędów topologicznych, a jednocześnie ujawniała wiele zależności spójnych z intuicją i wiedzą domenową (np. zależność jakości wina od zawartości alkoholu czy siarczanów).

8 Estymacja parametrów modelu

Po wybraniu najlepszego modelu (sieć HC), przystąpiono do etapu estymacji parametrów, czyli wyznaczenia warunkowych rozkładów prawdopodobieństwa (Conditional Probability Tables, CPT) dla każdego z węzłów w sieci. W przypadku zmiennych zależnych od innych (tj. posiadających rodziców), rozkład ten obliczany jest osobno dla każdej możliwej konfiguracji stanów rodziców.

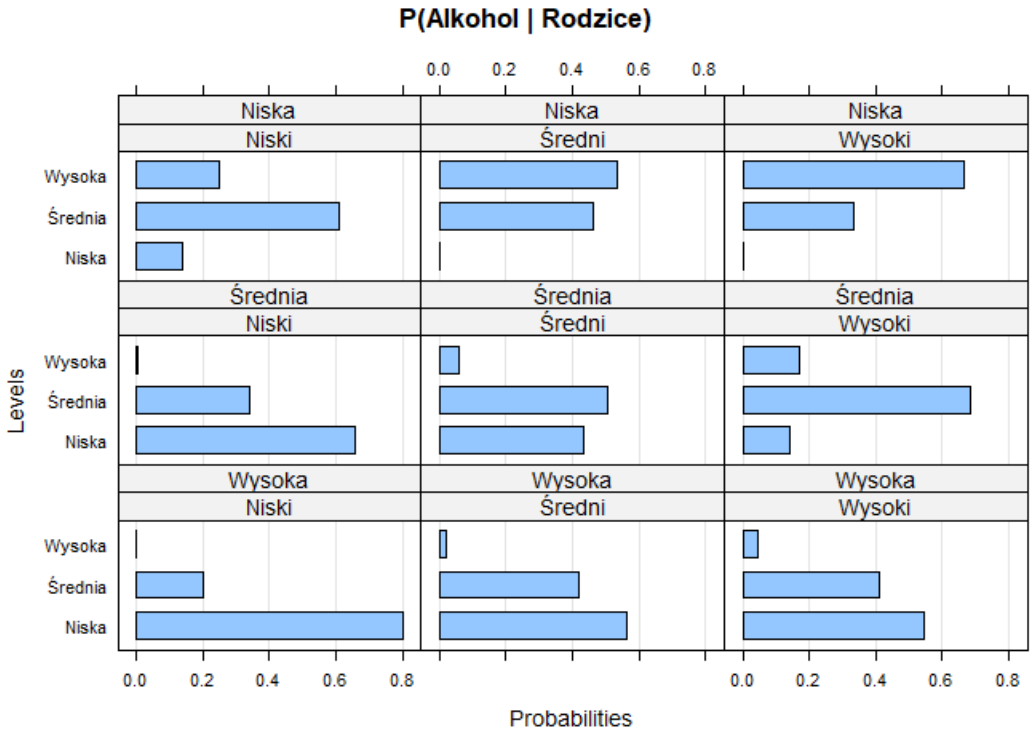
Wizualizacja wynikowego modelu HC z estymowanymi rozkładami węzłów przedstawiona została poniżej:



Rysunek 11: Sieć Bayesowska HC — Rozkłady Prawdopodobieństwa dla węzłów

Jednym z kluczowych elementów modelu jest zmienna wyjściowa `quality`, która zależy m.in. od zmiennej `alcohol`. Poniżej przedstawiono szczegółowy rozkład warunkowy $P(\text{quality} \mid \text{alcohol})$, tj. prawdopodobieństwa przyjęcia określonej jakości wina przy znanym poziomie zawartości alkoholu:

Poniższy wykres przedstawia warunkowy rozkład prawdopodobieństwa dla zmiennej `alcohol`, której wartości (Niska, Średnia, Wysoka) zależą od konfiguracji wartości zmiennych nadrzędnych (rodziców) w sieci Bayesowskiej. Każdy wykres w kratce reprezentuje rozkład prawdopodobieństwa dla jednej konkretnej kombinacji poziomów zmiennych rodzicielskich.

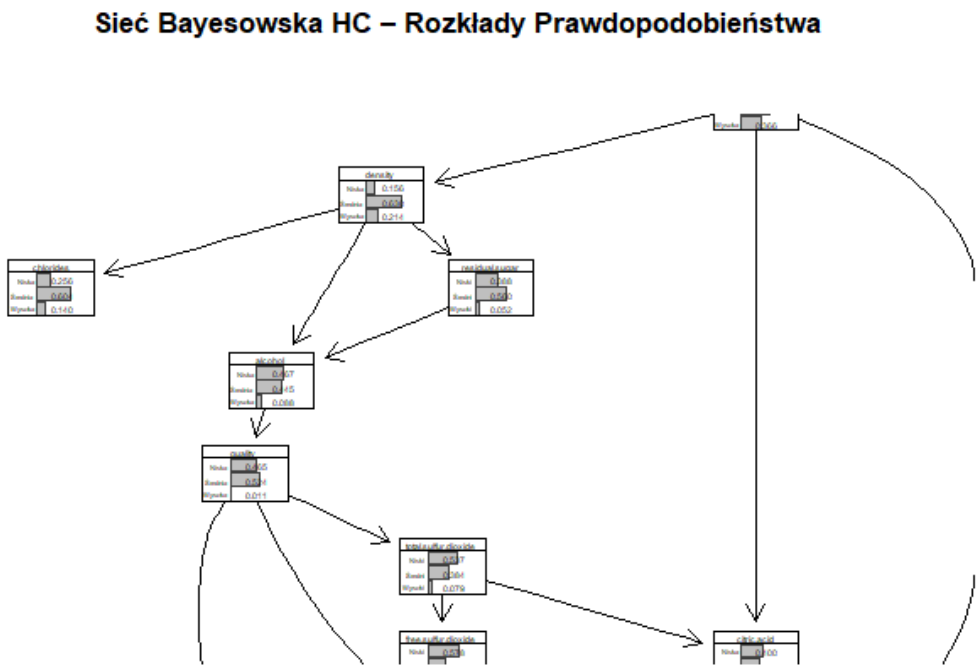


Rysunek 12: Rozkład $P($`alcohol` | Rodzice) — wpływ konfiguracji na prawdopodobieństwo zawartości alkoholu

Interpretacja:

- Dla wielu konfiguracji, widoczna jest przewaga wartości `Średnia` lub `Wysoka` alkoholu, co sugeruje dominację tych kategorii w zbiorze danych.
- Zmienność rozkładów w zależności od kombinacji rodziców wskazuje na istotny wpływ cech takich jak np. `density` lub `residual.sugar` (w zależności od konkretnej sieci), które występowały jako węzły nadrzędne.
- W niektórych przypadkach pojawiają się znaczne różnice — np. przy niskich wartościach rodziców, prawdopodobieństwo wysokiego alkoholu spada do zera.

Dzięki estymacji takiego rozkładu model może uwzględnić subtelne zależności między zmiennymi chemicznymi i ich wpływ na zawartość alkoholu, co jest kluczowe dla późniejszych predykcji jakości wina.



Rysunek 13: Rozkład $P(\text{quality} \mid \text{alcohol})$ — wpływ alkoholu na jakość wina

Interpretacja:

- Wina o **niskiej** zawartości alkoholu mają największe prawdopodobieństwo uzyskania niskiej jakości.
- Wina o **średniej** zawartości alkoholu najczęściej oceniane są jako średniej jakości.
- Wina o **wysokiej** zawartości alkoholu charakteryzują się wyraźnie wyższym prawdopodobieństwem osiągnięcia wysokiej jakości.

Takie wyniki są spójne z wiedzą ekspercką — alkohol wpływa nie tylko na smak, ale i na percepcję intensywności wina, co często znajduje odzwierciedlenie w ocenie sensorycznej.

Parametry wszystkich pozostałych węzłów (takich jak `density`, `citric.acid`, `chlorides` itd.) zostały również oszacowane w analogiczny sposób. Dzięki temu możliwe było przeprowadzenie dalszych analiz probabilistycznych i scenariuszy predykcyjnych.

9 Scenariusze predykcyjne

Na tym etapie przeprowadzono analizę scenariuszy predykcyjnych, czyli obliczenie prawdopodobieństw warunkowych różnych zmiennych wyjściowych w zależności od zadanych warunków (evidence) przy użyciu dopasowanej sieci Bayesowskiej HC.

Do obliczeń zastosowano funkcję `cpquery()` z pakietu `bnlearn`, która estymuje wartość warunkowego prawdopodobieństwa za pomocą próbkowania. Jako metodę wybrano `likelihood weighting` (LW), która działa efektywnie nawet w bardziej złożonych strukturach grafowych.

9.1 Scenariusz 1: $P(\texttt{fixed.acidity} = \textbf{Wysoka} \mid \texttt{density} = \textbf{Średnia})$

W tym scenariuszu obliczamy prawdopodobieństwo tego, że wino ma wysoką kwasowość stałą (`fixed.acidity = Wysoka`), pod warunkiem że jego gęstość (`density`) jest średnia. Formalnie oznacza to wyznaczenie wartości:

$$P(A = \text{Wysoka} \mid B = \text{Średnia}), \tag{1}$$

gdzie:

- $A =$ `fixed.acidity` — kwasowość stała (zmienna zależna),
- $B =$ `density` — gęstość wina (zmienna warunkująca).

Zgodnie z definicją prawdopodobieństwa warunkowego, korzystamy ze wzoru:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{2}$$

Ponieważ nie dysponujemy pełną tabelą częstości współwystępowania, estymację przeprowadzamy za pomocą próbkowania z ważeniem (likelihood weighting). Technika ta polega na:

- generowaniu wielu możliwych konfiguracji zmiennych zgodnie z rozkładem prawdopodobieństwa modelu (czyli z wcześniej wytrenowanej sieci Bayesowskiej),
- przypisaniu wag (likelihood) dla tych próbek, które spełniają warunki zadane w `evidence`,
- wyznaczeniu szacowanego prawdopodobieństwa jako liczby próbek spełniających zarówno warunek `density = Średnia`, jak i zdarzenie `fixed.acidity = Wysoka`, podzielonej przez wszystkie próbki, które spełniały tylko warunek dowodu.

Kod w R:

```
scenariusz1 <- cpquery(fit_hc,
                       event = (fixed.acidity == "Wysoka"),
                       evidence = list(density = "Średnia"),
                       method = "lw")
```

Uzyskany wynik:

$$P(\texttt{fixed.acidity = Wysoka} \mid \texttt{density = Średnia}) \approx \boxed{0.0903}$$

Interpretacja:

- Gdy wino ma średnią gęstość, istnieje około 9% szans, że jego kwasowość stała będzie wysoka.
- Taki wynik wskazuje na dość niską korelację bezpośrednią, ale może mieć znaczenie w połączeniu z innymi zmiennymi.

```
scenariusz4 <- cpquery(fit_hc,
                       event = (fixed.acidity == "Wysoka"),
                       evidence = list(density = "Srednia"),
                       method = "lw")

cat("Scenariusz 4 - Prawdopodobieństwo Warunkowe wysokiej kwasowości stałej przy średniej gęstości:", scenariusz4, "\n")
```

Rysunek 14: Scenariusz 1: Estymacja $P(\texttt{fixed.acidity = Wysoka} \mid \texttt{density = Średnia})$

9.2 Scenariusz 2: $P(\texttt{quality} = \textbf{Wysoka} \mid \texttt{alcohol} = \textbf{Wysoka}, \texttt{sulphates} = \textbf{Wysoka})$

Sprawdzamy, jak prawdopodobne jest, że wino ma wysoką jakość, gdy zawartość alkoholu oraz siarczanów jest wysoka:

```
scenariusz2 <- cpquery(fit_hc,
                       event = (quality == "Wysoka"),
                       evidence = list(alcohol = "Wysoka",
                                       sulphates = "Wysoka"),
                       method = "lw")
```

Wynik: **0.1512**

```
> scenariusz1 <- cpquery(fit_hc,
+                        event = (quality == "Wysoka"),
+                        evidence = list(alcohol = "Wysoka", sulphates = "Wysoka"),
+                        method = "lw")
> cat("Scenariusz 1 - Prawdopodobieństwo wysokiej jakości wina przy wysokim alkoholu i siarczynach:", scenariusz1, "\n")
Scenariusz 1 - Prawdopodobieństwo wysokiej jakości wina przy wysokim alkoholu i siarczynach: 0.1521245
```

Rysunek 15: Scenariusz 2: $P(\texttt{quality = Wysoka} \mid \texttt{alcohol, sulphates} = \text{Wysoka})$

9.3 Scenariusz 3: $P(\texttt{quality} = \textbf{Wysoka} \mid \texttt{volatile.acidity} = \textbf{Niska}, \texttt{residual.sugar} = \textbf{Średni})$

```
scenariusz3 <- cpquery(fit_hc,
                       event = (quality == "Wysoka"),
                       evidence = list(volatile.acidity = "Niska",
                                       residual.sugar = "Średni"),
                       method = "lw")
```

Wynik: **0.0062**

```
> scenariusz2 <- cpquery(fit_hc,
+                        event = (quality == "Wysoka"),
+                        evidence = list(volatile.acidity = "Niska", residual.sugar = "Sredni"),
+                        method = "lw")
> cat("Scenariusz 2 - Prawdopodobieństwo wysokiej jakości wina przy niskiej kwasowości i średnim cukrze:", scenariusz2, "\n")
Scenariusz 2 - Prawdopodobieństwo wysokiej jakości wina przy niskiej kwasowości i średnim cukrze: 0.006247147
```

Rysunek 16: Scenariusz 3: Wpływ niskiej kwasowości lotnej i średniego cukru resztkowego

9.4 Scenariusz 4: $P(\texttt{quality} = \textbf{Wysoka} \mid \texttt{density} = \textbf{Średnia}, \texttt{pH} = \textbf{Średnie}, \texttt{alcohol} = \textbf{Niska})$

```
scenariusz4 <- cpquery(fit_hc,
                       event = (quality == "Wysoka"),
                       evidence = list(density = "Średnia",
                                       pH = "Średnie",
                                       alcohol = "Niska"),
                       method = "lw")
```

Wynik: **0.0027**
Interpretacja:

- Tak niskie prawdopodobieństwo oznacza, że połączenie niskiej zawartości alkoholu z neutralnym pH i średnią gęstością nie sprzyja wysokiej jakości wina.
- Wynik ten potwierdza wcześniejsze obserwacje, że alkohol i siarczany są silnymi predyktorami jakości.

```
> scenariusz3 <- cpquery(fit_hc,
+                        event = (quality == "Wysoka"),
+                        evidence = list(density = "Srednia", pH = "Srednie", alcohol = "Niska"),
+                        method = "lw")
> cat("Scenariusz 3 - Prawdopodobieństwo wysokiej jakości wina przy średniej gęstości, pH i niskim alkoholu:", scenariusz3, "\n")
Scenariusz 3 - Prawdopodobieństwo wysokiej jakości wina przy średniej gęstości, pH i niskim alkoholu: 0.002709811
```

Rysunek 17: Scenariusz 4: Wpływ niskiego alkoholu, neutralnego pH i średniej gęstości

10 Podsumowanie

Celem niniejszego projektu było zbadanie struktury zależności między zmiennymi chemicznymi opisującymi czerwone wina portugalskie na podstawie zbioru danych *Wine Quality Dataset* dostępnego na platformie Kaggle. W tym celu zastosowano sieci Bayesowskie jako narzędzie do modelowania probabilistycznych relacji między zmiennymi oraz do predykcji jakości wina.

Projekt przebiegał wieloetapowo:

- **Wczytano i przygotowano dane**: przekształcono wszystkie zmienne ilościowe do postaci kategorycznej poprzez dyskretyzację, co umożliwiło ich wykorzystanie w modelach Bayesowskich.
- **Przeprowadzono testy istotności** χ^2: dzięki czemu uzyskano wgląd w najsilniejsze zależności pomiędzy parami zmiennych.
- **Zbudowano dziewięć sieci Bayesowskich**: wykorzystano metody takie jak HC, Tabu, PC, GS, IAMB i jego warianty. Każda z sieci została poddana analizie struktury, a w przypadku niespójności – ręcznej poprawie.
- **Dokonano porównania modeli na podstawie wartości BIC**: najlepszy rezultat osiągnął model Hill Climbing (HC), co uzasadniło jego wybór do dalszych analiz.
- **Przeprowadzono estymację parametrów modelu HC**: umożliwiając wizualizację rozkładów warunkowych dla wybranych węzłów, takich jak `quality` i `alcohol`.
- **Zbudowano i przeanalizowano scenariusze predykcyjne**: m.in. prawdopodobieństwo wysokiej jakości wina przy wysokim alkoholu i siarczanach, czy też wysoka kwasowość stała przy średniej gęstości.

Zwieńczeniem projektu było przeprowadzenie dokładnej analizy jednego ze scenariuszy z rozpisaniem wzorów oraz interpretacją wyniku.

Całość prac pokazała, że sieci Bayesowskie są skutecznym i przejrzystym narzędziem do:

- eksploracji relacji między cechami chemicznymi produktu,
- predykcji trudnych do zmierzenia atrybutów (jak jakość),
- tworzenia scenariuszy decyzyjnych i wspomagania podejmowania decyzji.

Z uwagi na wysoką jakość dopasowania (najniższy BIC dla modelu HC) oraz sensowność logiczną uzyskanych połączeń między zmiennymi, model ten można uznać za solidną podstawę do dalszych analiz i zastosowań np. w kontroli jakości produkcji wina.