Probabilistyczne uczenie maszynowe
Projekt 1

Michał Maternik, Maksym Telepchuk

14 maja 2020

Streszczenie

Zadanie polegało na przeprowadzeniu eksploracyjnej analizy wybranego zbioru danych oraz implementacji dwóch modeli probabilistycznych, w tym graficznego modelu probabilistycznego. Zostało ono zrealizowane dla zbioru dotyczącego grzybów, na podstawie którego przewidujemy czy grzyb jest jadalny czy nie.

1 Eksploracyjna analiza danych

Wybrany model danych dotyczy grzybów, które opisane są 23 cechami wyszczególnionymi poniżej. Ten zestaw danych zawiera opisy hipotetycznych próbek odpowiadające 23 gatunkom grzybów. Każdy gatunek jest identyfikowany jako zdecydowanie jadalne, zdecydowanie trujące lub o nieznanej jadalności. Ta ostatnia klasa została połączona z trującymi w jedną. Eksperty wyraźnie stwierdzają, że nie ma prostej zasady określania jadalności grzyba, jednak są cechy grzybów, które są bardziej lub mniej skorelowane między sobą.

Poniżej znajdują się statystyki cech.

	Column	Non-Null Count	Dtype
0	class	8124 non-null	category
1	cap-shape	8124 non-null	category
2	cap-surface	8124 non-null	category
3	cap-color	8124 non-null	category
4	bruises?	8124 non-null	category
5	odor	8124 non-null	category
6	gill-attachment	8124 non-null	category
7	gill-spacing	8124 non-null	category
8	gill-size	8124 non-null	category
9	gill-color	8124 non-null	category
10	stalk-shape	8124 non-null	category
11	stalk-root	5644 non-null	object
12	stalk-surface-above-ring	8124 non-null	category
13	stalk-surface-below-ring	8124 non-null	category
14	stalk-color-above-ring	8124 non-null	category
15	stalk-color-below-ring	8124 non-null	category
16	veil-type	8124 non-null	category
17	veil-color	8124 non-null	category
18	ring-number	8124 non-null	category
19	ring-type	8124 non-null	category
20	spore-print-color	8124 non-null	category
21	population	8124 non-null	category
22	habitat	8124 non-null	category

Liczebność zbioru wynosi 8124 egzemplarze. Wszystkie kolumny mają charakter kategoryczny.

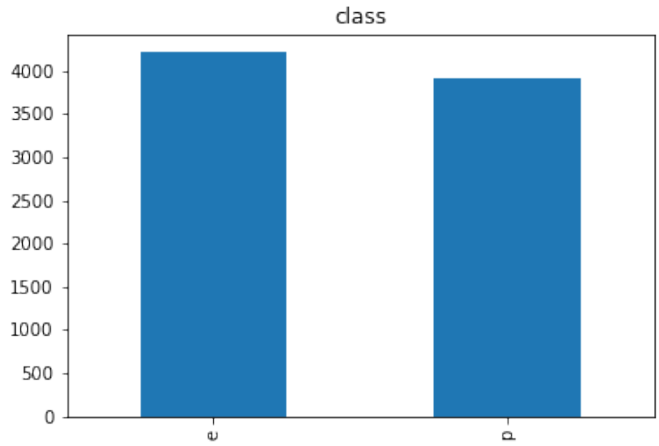
Znaczenie dostępnych kolumn jest następujące:

- class - jadalność (jadalny, niejadalny)

- cap-shape - kształt kapelusza (dzwon, płaski, wcięty)
- cap-surface - powierzchnia kapelusza (włóknista, łuszcząca się, gładka)
- cap-color - kolor kapelusza (żółty, biały...)
- bruises? - czy grzyb jest siniakiem? (tak / nie)
- odor - zapach: (migdałowy, ostry, pikantny, nijaki)
- gill-attachment - sposób przyrastania blaszek do trzonu (przyrośnięte, wolne)
- gill-spacing - odstęp między blaszkami (zbliżone, oddalone)
- gill-size - szerokość blaszki (szeroki, wąski)
- gill-color - kolor blaszki (czarny, brązowy, biały)
- stalk-shape - kształt trzonu (pogrubiający się, zwężający się)
- stalk-root - korzeń trzonu: (bulwiasty, stożkowy)
- stalk-surface-above-ring - typ powierzchni trzonu nad pierścieniem (włóknista, łuskowata, gładka)
- stalk-surface-below-ring -typ powierzchni trzonu poniżej pierścienia(włóknista, łuskowata, gładka)
- stalk-color-above-ring - kolor trzonu nad pierścieniem (brązowy, szary ...)
- stalk-color-below-ring - kolor trzonu poniżej pierścienia (brązowy, szary ...)
- veil-type - typ osłony grzyba (częściowa, całkowita)
- veil-color - kolor osłony grzyba (brązowy, biały . . .)
- ring-number - liczba pierścieni (brak, jeden, dwa)
- ring-type - typ pierścienia (z pajęczyną, zanikający, rozszerzający się, duży, brak, zwisający, ślad pierścienia)
- spore-print-color - kolor wysypu zarodników: (czarny, brązowy, buff, czekoladowy, zielony, pomarańczowy, fioletowy, biały, żółty)
- population - typ populacji (obfita, skupiona, liczna, rozproszona, rzadka, samotna)

- habitat - siedlisko (trawy, liście, łąki, ścieżki, urban, odpady, lasy)

Kolumna class, która oznacza, czy grzyb jest jadalny, będzie stanowiła wyznaczaną w modelach wartość.



Rysunek 1: Kolumna Class

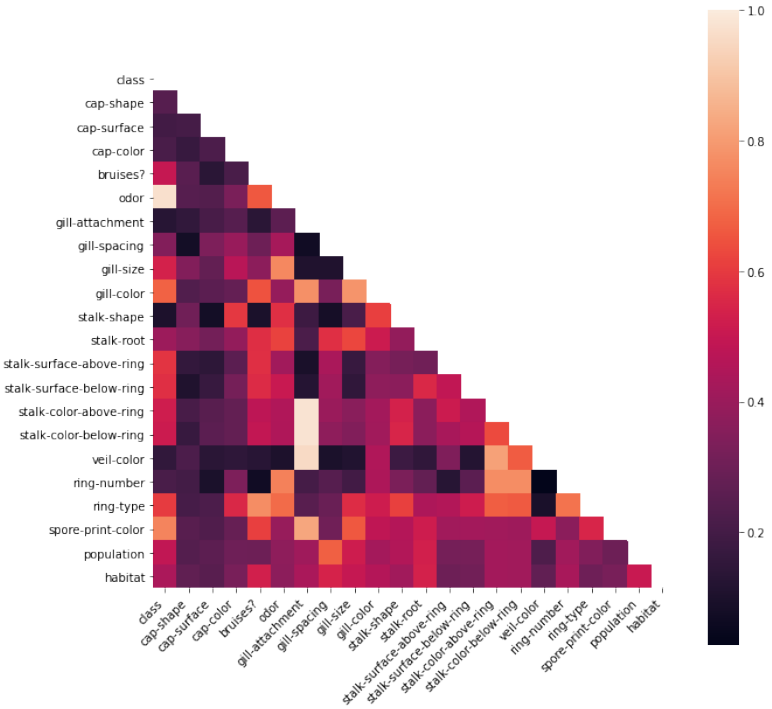
Z jej rozkładu na rysunku 1 widzimy, że zbiór jest pod jej względem dobrze zrównoważony i jego dodatkowe balansowanie przy wykorzystaniu jednej z wielu dostępnych metod nie było konieczne.

Zaobserwowano, że spośród pozostałych cech kolumna 'stalk-root' posiadała brakujące wartości i dlatego została usunięta. Sugerowane jest użycie mechanizmów, takie jak maski, które uwzględniają wartości brakujące.

Zauważono także, że kolumna 'veil-type' posiadała tylko jedną wartość, stąd byłaby więc przydatna przy klasyfikacji i ze względu na to również została usunięta.

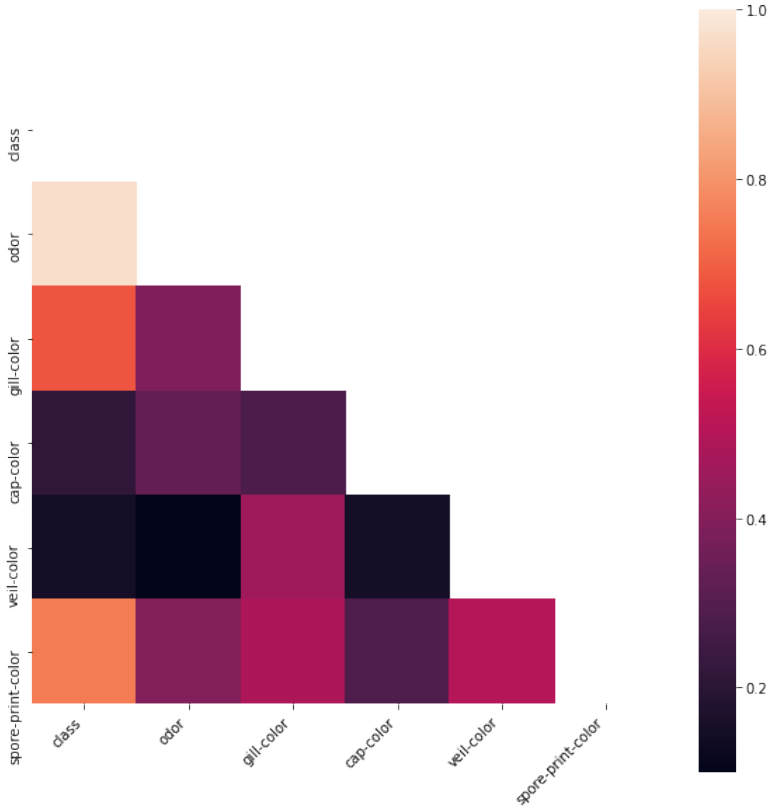
Ponieważ nie można stosować zwykłej korelacji pomiędzy atrybutami kategorycznymi, do wyznaczenie wartości, które będą wyglądały jak korelacja, ale będą działać z wartościami kategorycznymi, została użyta metoda Cramér's V. Opiera się ona na teście chi-kwadrat Pearsona. Jest ona symetryczna oraz wartości są z zakresu $[0, 1]$

Na rys. 3 są przedstawione wyniki przeprowadzenia miary tego typu korelacji dla każdej z kategorii.

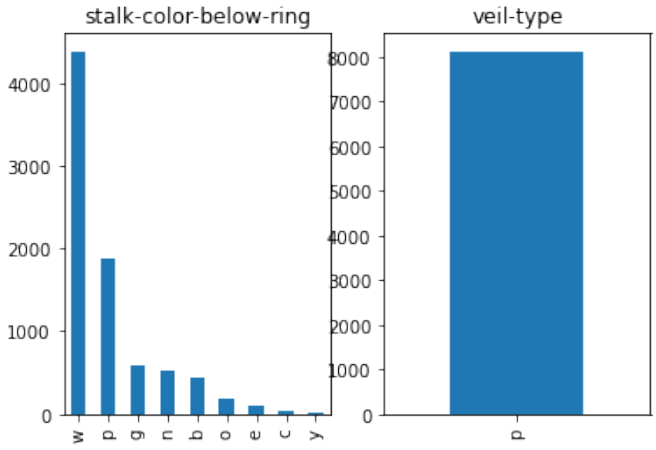


Rysunek 2: Podobieństwo pomiędzy cechami.

Z wykresu można odczytać to, że dużo cech zależy od koloru blaszki, a jadalność grzyba jest szczególnie mocno skorelowana z kolorem różnych składowych oraz z zapachem grzyba.



Rysunek 3: Podobieństwo pomiędzy cechami koloru, jadalności i zapachu.



Rysunek 4: Kolumny 'stalk-color-below-ring' oraz 'veil-type'

W przypadkach innych cech nie wystąpiły wartości brakujące.

2 Modele

2.1 Naive Bayes

Wybranym prostym modelem probabilistycznym był Naive Bayes. Fundamentalnym założeniem tego modelu jest założenie, że poszczególne kolumny lub cechy są od siebie niezależne. Założenie to często nie jest w pełni spełnione i stąd pochodzi pierwszy człon jego nazwy (naiwny). Analiza naszych danych doprowadziła do wniosku, że i w tym przypadku założenie to nie dla wszystkich par kolumn będzie spełnione, co niewątpliwie miało wpływ na uzyskiwane wyniki.

W modelu Naiwnego Bayesu wykorzystywana jest zależność między prawdopodobieństwem warunkowym przynależności egzemplarza do klasy, a iloczynem prawdopodobieństw warunkowych wystąpienia poszczególnych cech warunkowanych przynależnością do tej klasy oraz całkowitego prawdopodobieństwa przynależności do danej klasy. Tak określone prawodpodobieństwo posterior wynikające z wnioskowania Bayesowskiego określa relacja proporcjanolności:

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

W implementacji zamiast iloczynu wykorzystujemy sumę logarytmów ze względu na to, że eliminuje to sytuację, gdy z danych treningowych wynikałoby zerowe prawdopodobieństwo wystąpienia którejś z cech, która przez to zerowała by nam cały iloczyn. W rzeczywistości oczekujemy, że nawet jeśli takie prawdopodobieństwo jest bardzo małe, to nie jest zerowe. Zastosowanie sumy logarytmów nie wpływa przy tym na poprawność przeprowadzanego wnioskowania.

Klasa docelowa (grzyb jadalny lub nie) określana jest poprzez wybranie klasy najbardziej prawdopodobnej:

$$y = \operatorname*{argmax}_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Jeśli chodzi o realizację, to ponadto wykorzystana została implementacja w pyro, wykorzystująca mechanizm pyro.infer.SVI oraz TraceEnum_ELBO. Natomiast predykcja wykonywana została przy wykorzystaniu pyro.infer.Predictive.

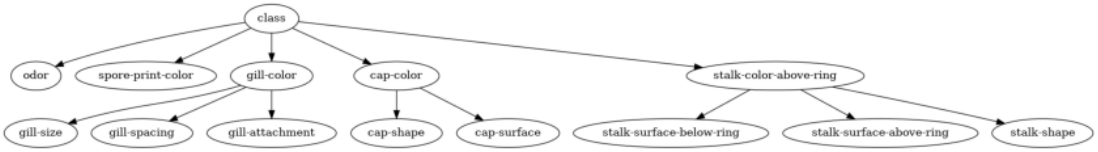
Zastosowana implementacja obsługuje wyłącznie kolumny kategoryczne, ponieważ tylko takie występowały w wykorzystywanym zbiorze danych; była więc wystarczająca dla naszych potrzeb.

2.2 Probabilistic Graphical Model - Bayesian Network

Spośród probabilistycznych modeli graficznych zastosowany został skierowany model sieci Bayesowskiej.

Poznanie struktury sieci bayesowskich może być skomplikowane z dwóch głównych powodów : trudności w wnioskowaniu o przyczynowości i bardzo dużej liczby możliwych skierowanych krawędzi, które mogłyby istnieć w zbiorze danych. Problem pojawia się, gdy algorytm uczenia struktury bierze pod uwagę tylko korelację lub inną miarę współwystępowania w celu ustalenia, czy krawędź powinna istnieć. W części eksploracji danych zostało pokazane, że nie zawsze z tego da się skutecznie takie zależności wywnioskować.

Po rozmowie z ekspertem w dziedzinie grzybów, został sformułowany model, który zdaniem eksperta może być skuteczny w klasyfikacji jadalności grzyba. Zależności pomiędzy cechami zaproponowane przez eksperta są podane na rys. 5



Rysunek 5: Model zaproponowany przez eksperta

Naiwnym podejściem w przypadku znajdywania takich zależności byłoby przeszukiwanie wszystkich możliwych ukierunkowanych grafów acyklicznych i identyfikowanie tego grafu, który minimalizuje funkcje celu. Naiwna implementacja tego wyszukiwania ma złożoność obliczeniową nadwykładniczą i staje się niemożliwa jeśli weźmie się pod uwagę nawet kilkanaście zmiennych. Jednak programowanie dynamiczne może skutecznie usunąć wiele powtarzanych obliczeń i zredukować je do wykładniczego czasu.

Biblioteka **pomegranate** daje możliwość przeprowadzenie takiego wyszukiwania za pomocą algorytmu A* oraz programowania dynamicznego, jednak w przypadku tego zbioru danych to jest bardzo czasochłonne, wyszukiwanie struktury zajmuje godziny czasu.

Następnym typem wyszukiwania struktury jest podejście zachłanne. Metoda zachłannego wyszukiwania iteracyjnie znajduje najlepszą zmienną, którą można dodać do rosnąco posortowanego porządku topologicznego, umożliwiając nowej zmiennej rysowanie tylko ze zmiennych znajdujących się już w porządku topolo-

gicznym. Ma dobrą równowagę między tworzeniem dobrych (często optymalnych) wykresów a niewielkim kosztem obliczeniowym i powierzchnią pamięci. Nie ma jednak gwarancji, że dzięki temu powstanie optymalny globalnie wykres.

Jednak nawet takie względnie szybkie podejście jest nie odpowiednie w przypadku tego zbioru danych. Algorytm znajduje dużo połączeń pomiędzy zmiennymi, co powoduje dużą ilość parametrów, które się nie mieszczą w pamięci komputerowej

Finalnie, aby uzyskać dobre struktury w rozsądnym czasie został użyty algorytm budowy drzewa Chow-Liu [1], która uczy się optymalnego drzewa na podstawie danych. Zasadniczo algorytm oblicza wzajemną informację między wszystkimi parami zmiennych, a następnie znajduje maksymalne drzewo rozpinające. Algorytmem jest o złożoności $O(d^2)$ (d - liczba atrybutów) i praktycznie jest szybki i efektywny pod względem pamięci, choć tworzy struktury o gorszym prawdopodobieństwie warunkowym.

Model sieci Bayesowskich został zaimplementowany za pomocą pyro w klasie PGM. Do tej klasy jest przekazywany graf skierowany (musi być być acykliczny) z zależnościami między zmiennymi.

Przy uczeniu model rozważamy obserwacje x, zmienną ukrytą z i parametry θ. Gęstość prawdopodobieństwa ma postać

$$p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$$

.

W kontekście wyboru najlepszego modelu musi być wybrany taki parametr θ, że

$$\theta_{max} = \underset{\theta}{\operatorname{argmax}} \log p_\theta(x)$$

Na podstawie przekazywanego grafu połączeń są inicjalizowane parametry w pyro, które są uczone iteracyjnie przy pomocy SVI. Czyli do każdej zmiennej jest przypisana macierz parametrów, w której rozkład prawdopodobieństwa dla tej zmiennej jest wybierany na podstawie wyników próbkowania zmiennych „rodzicowych".

Podczas trenowania wszystkie zmienne są obserwowane, natomiast podczas testowania, cecha klasy nie jest obserwowana oraz jest wyliczana za pomocą klasy Predictive na podstawie wytrenowanych parametrów.

W takim układzie obliczenie $\log p_\theta(x)$ jest zadaniem trywialnym.

3 Eksperymenty

Scenariusz testowy został przygotowany w ten sposób, że 80 procent danych przeznaczonych zostało do przeprowadzenia uczenia, a pozostałe 20 procent do testów uzyskanego modelu.

W celu wykonania oceny poszczególnych modeli wykorzystane zostały metryki accuracy, precision, recall oraz F1-Score. Opierają się one na macierzy pomyłek, która w przypadku klasyfikacji binarnej składa się z 4 wartości:

- TP - True Positive - ilość poprawnie rozpoznanych egzemplarzy danej klasy
- FP - False Positive - ilość egzemplarzy błędnie rozpoznanych jako przynależne do klasy
- FN - False Negative - ilość egzemplarzy błędnie rozpoznanych jako nieprzynależące do danej klasy
- TN - True Negative - ilość egzemplarzy poprawnie rozpoznanych jako nieprzynależące do danej klasy

Na ich podstawie obliczane są zastosowane metryki:

- $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
- $Precision = \frac{TP}{TP+FP}$
- $Recall \frac{TP}{TP+FN}$
- $F1 - Score = \frac{2*Recall*Precision}{Recall+Precision}$

Accuracy określa więc proporcję poprawnie rozpoznanych klas do wszystkich poddanych ocenie egzemplarzy, Precision proporcję poprawnie rozpoznanych egzemplarzy klasy do wszystkich rozpoznanych jako do niej przynależne, a Recall proporcję poprawnie rozpoznanych egzemplarzy danej klasy do wszystkich, które faktycznie do niej przynależą. Ponieważ klasyfikujemy przydatność grzyba do spożycia, praktyczne wykorzystanie wymagałoby bardzo wysokich wartości wszystkich parametrów, a szczególnie wysokiej precyzji w klasyfikowaniu jadalności.

Każda z nich liczona było pięciokrotnie w procesie stratyfikowanej krosswalidacji, a następnie uzyskane wyniki zostały uśrednione. Zbadane zostało także odchylenie standardowe metryk dla poszczególnych modeli.

3.1 Naive Bayes

Algorytm Naive Bayes sprawdził się w tym przypadku bardzo dobrze uzyskując wyniki poszczególnych miar na poziomie 95 procent. Również odchylanie standardowe uzyskane dla wyników w procesie kroswalidacji pozostało względnie niewielkie.

Metryka	Wartość średnia	Odchylenie standardowe
accuracy	0.8816	0.0076
precision	0.9564	0.0046
recall	0.7903	0.0133
fscore	0.8654	0.0094

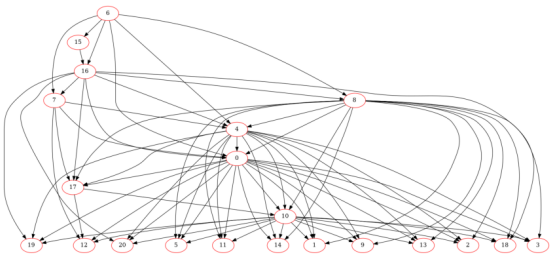
Wyniki wskazują że uzyskany model dobrze klasyfikuje przynależność grzyba do danej klasy, ponieważ jednak chodzi tu o przydatność wskazanego grzyba do spożycia, to wartość metryki precision na poziomie 95 procent może okazać się zbyt niska do praktycznego wykorzystania - ryzyko pomyłki wydaje się w tym przypadku zbyt duże.

3.2 Bayesian Network

Zależności pomiędzy atrybutami zostały wyliczone za pomocą biblioteki **pomegranate**.

3.2.1 Podejście zachłanne

Na początku zostało użyte podejście zachłanne, które wygenerowało strukturę zwizualizowaną na rysunku 6.



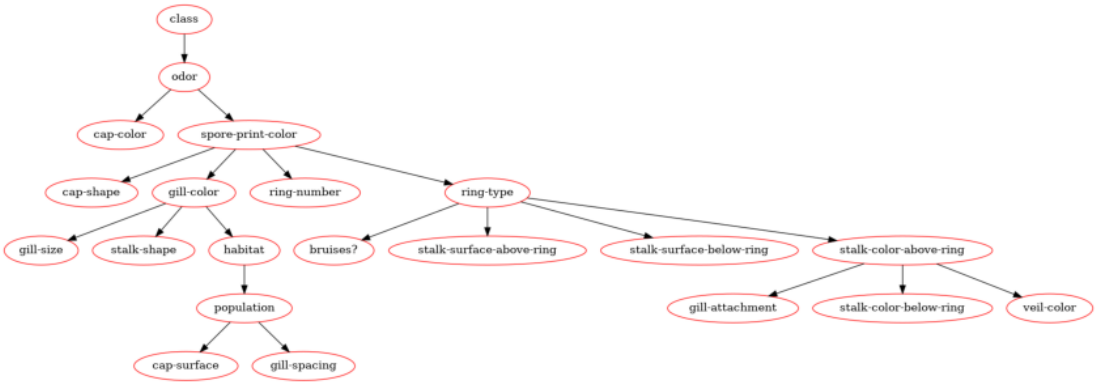
Rysunek 6: Zachłanne podejście

Policzenie tej struktury zajęło 25 min, oraz model nie nadaję się na uczenie ze względu na to, że parametry dla takiego modelu nie mieszczą się w posiadanej

pamięci komputerowej.

3.2.2 Podejście aproksymacyjne

W tym eksperymencie został użyty algorytm Chow Liu. Policzenie struktury modelu zajmuje od 2 do 10 sekund. Zależności, które są pozyskiwane, są rzadkie i łatwo czytelne. Wizualizacja struktury jest podana na rys. 7.



Rysunek 7: Struktura uzyskana za pomocą algorytmu Chow Liu

Metryka	Wartość średnia	Odchylenie standardowe
accuracy	0.9346	0.0272
precision	0.9545	0.0029
recall	0.9078	0.0614
fscore	0.9294	0.0318

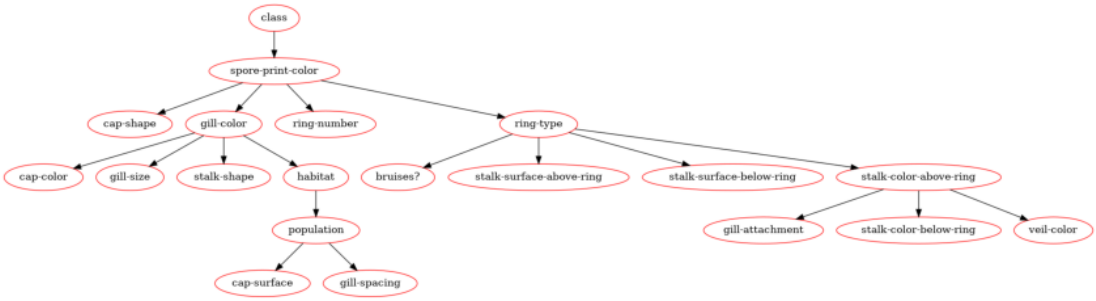
Metryka czułości względem niejadalnych grzybów jest najważniejszą ze stosowanych, ponieważ w zdefiniowanym problemie jej niska wartość stanowi największe zagrożenie dla użytkownika.

Z wartości metryk wynika, że algorytm dobrze potrafi klasyfikować grzyby jadalne i niejadalne. Niemniej jednak, wartość czułości względem grzybów jadalnych jest wyższa niż dla grzybów niejadalnych, co jest bardziej niebezpieczne niż w sytuacji odwrotnej.

3.2.3 Podejście aproksymacyjne dla modelu bez atrybutu kluczowego

Z powyższego wynika, że jeden z atrybutów jest kluczowy, czyli taki, na podstawie którego można bardzo precyzyjnie wyznaczyć jadalność grzyba. Jeżeli policzyć tym samym algorytmem zależności bez tego atrybutu, to atrybut spore-print-color jest uznawany za algorytm jako jedyny, od którego zależy klasa jadalności. Przy

takim uproszczeniu taki model będzie miał gorsze wyniki, jednak wykazuje to, że algorytm wykrył tylko najbardziej istotne zależności i taki model jest tylko przybliżeniem rzeczywistości.



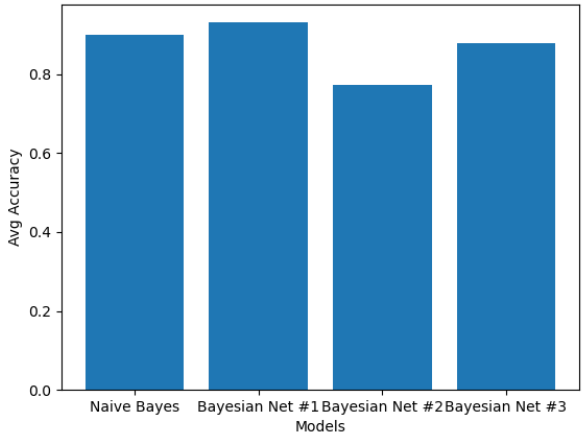
Rysunek 8: Model uzyskany przez algorytm Chow Liu bez cechy zapachu

Metryka	Wartość średnia	Odchylenie standardowe
accuracy	0.7742	0.1303
precision	0.8198	0.1483
recall	0.7741	0.2022
fscore	0.7648	0.1253

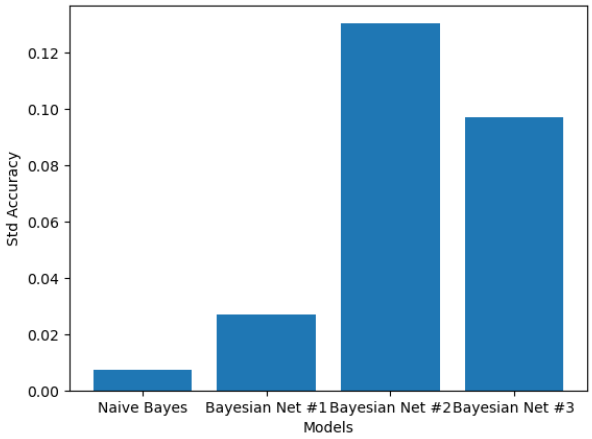
Można wywnioskować, że bez cechy zapachu czułość dla grzybów niejadalnych obniża się o 15%. Wykazuje to, że bez tej cechy algorytm istotnie traci na skuteczności klasyfikacji jadalności grzyba.

3.3 Porównanie modeli

Każdy z modeli został przetestowany ujednoliconym procesem polegającym na wykonaniu stratyfikowanej pięciokrotnej kroswalidacji. Wartości metryk zostały zebrane i uśrednione, a następnie zwizualizowane na diagramach.

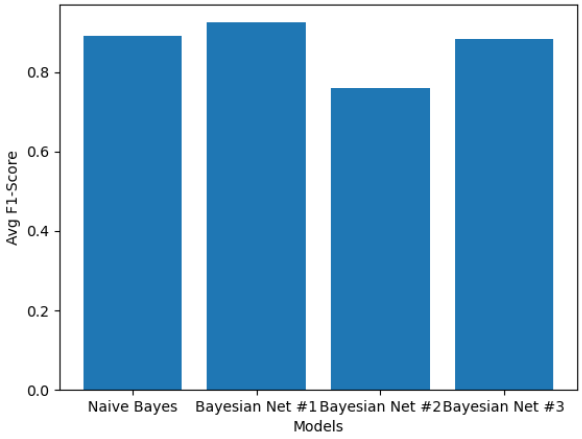


Rysunek 9: Porównanie metryk Accuracy dla poszczególnych modeli

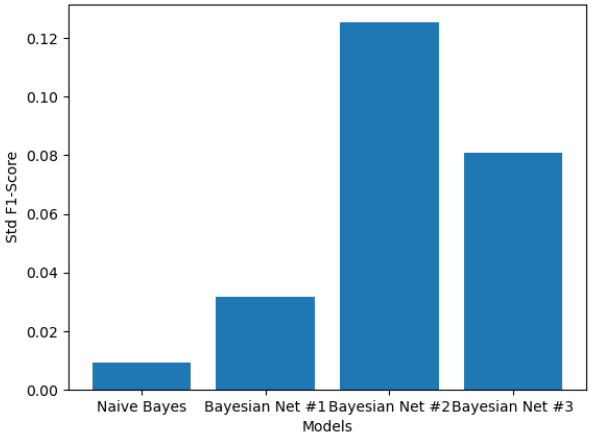


Rysunek 10: Porównanie odchyleń standardowych metryki Accuracy dla poszczególnych modeli

Pod względem dokładności najlepiej wypadła sieć uzyskana za pomocą pomegrenade, za nią uplasował się algorytm Naive Bayes. Pozostałe sieci, w tym uzyskana na podstawie wskazówek od naszego eksperta nie były aż tak dobre.

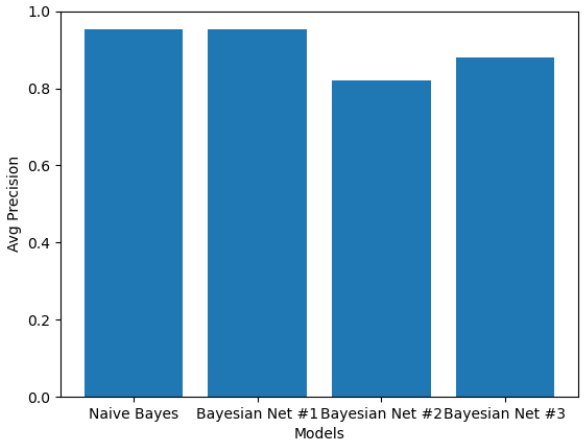


Rysunek 11: Porównanie metryk F1-Score dla poszczególnych modeli

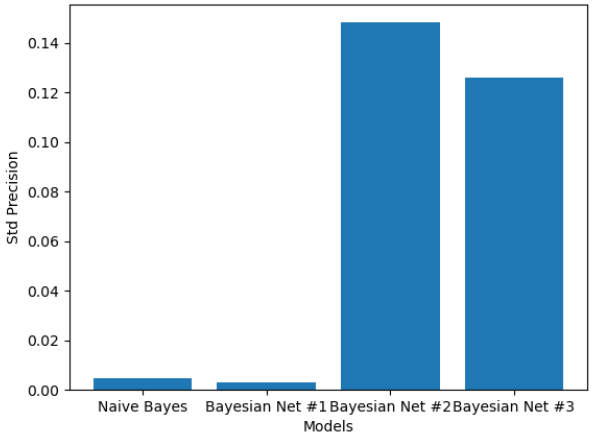


Rysunek 12: Porównanie odchyleń standardowych metryki F1-Score dla poszczególnych modeli

Porównanie za pomocą metryki F1-score wypada podobnie jak dla dokładności. Ponownie najlepsza okazała się sieć uzyskana za pomocą pomegranade.

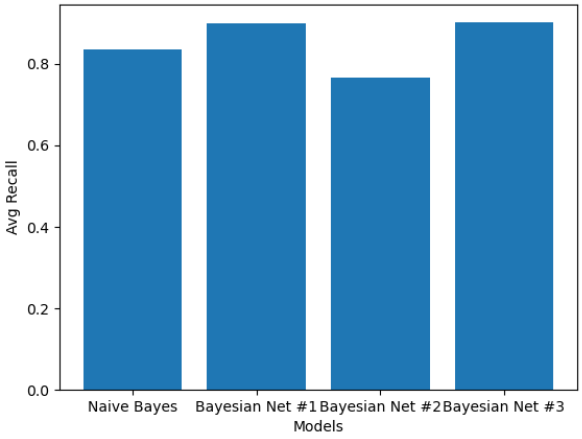


Rysunek 13: Porównanie metryk Precision dla poszczególnych modeli

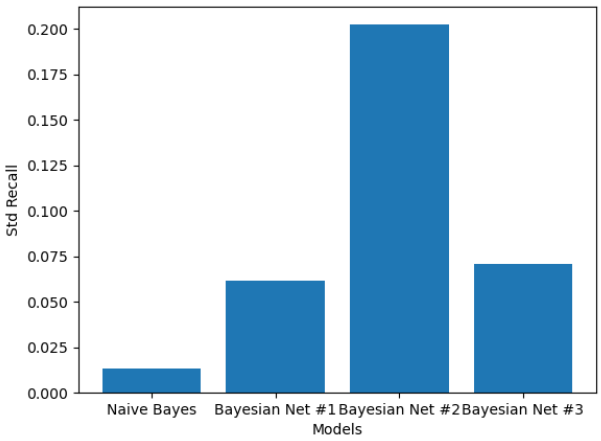


Rysunek 14: Porównanie odchyleń standardowych metryki Precision dla poszczególnych modeli

Pod względem precyzji nasze dwa pierwsze modele uzyskały bardzo zbliżone wyniki.



Rysunek 15: Porównanie metryk Recall dla poszczególnych modeli



Rysunek 16: Porównanie odchyleń standardowych metryki Recall dla poszczególnych modeli

Przy porównaniu pod względem metryki Recall zauważamy, że w tym przypadku niedoceniana dotychczas sieć zbudowana zgodnie ze wskazówkami naszego eksperta uzyskuje wyniki lepsze od modelu naiwnego Bayesa i zbliżone do sieci uzyskanej za pomocą pomegrenade.

4 Wnioski

Podsumowując możemy uznać, że dwa z uzyskanych modeli osiągnęły bardzo dobre wyniki na poziomie 90 procent, co w połączeniu z niskim odchyleniem standardowym stanowi o ich przydatności do rozpoznawania przynależności grzybów do odpowiedniej klasy jadalności. Był to model Naiwnego Bayesa oraz pierwszy rozważany model graficzny, który okazał się najlepszy. Pozostałe dwa modele okazały się gorsze pod względem średnich uzyskanych metryk, ale przede wszystkim o wiele większych odchyleniach standardowych, co wskazuje na dużą ich niestabilność i zależność od wykorzystanego zbioru danych uczących.

Najlepszy model udało nam się uzyskać wykorzystując probailistyczny model graficzny zbudowany według grafu zasugerowanego przez bibliotekę pomegrenade za pomocą algorytmu 'chow-liu'. Jest to najprostszy z oferowanych algorytmów, ale mimo to okazał się w tym przypadku bardzo skuteczny.

Jeśli chodzi o przygotowanie danych wejściowych dla modeli, to w retrospekcji zauważamy, że zamiast usuwać kolumnę z brakującymi wartościami można było przyjąć inne rozwiązanie polegające na zastosowaniu maskowania brakującej wartości za pomocą poutine.mask. Takie podejście zastosujemy prawdopodobnie w kolejnych budowanych modelach.

Literatura

[1] https://en.wikipedia.org/wiki/Chow%E2%80%93Liu_tree.

[2] https://archive.ics.uci.edu/ml/datasets/Mushroom.