1. Krótki opis analizowanego zbioru danych

Zbiór danych jest wyczyszczoną wersją oryginalnego zbioru danych dotyczących grzybów do klasyfikacji binarnej dostępnego w bibliotece UCI. Zbiór danych został wyczyszczony za pomocą różnych technik, takich jak imputacja modalna, kodowanie one–hot, normalizacja Z-score i selekcja cech. Zawiera 9 kolumn:

- Średnica kapelusza [Cap Diameter]
- Kształt kapelusza [Cap Shape]
- Przyczepność blaszek [Gill Attachment]
- Kolor blaszek [Gill Color]
- Wysokość trzonu [Stem Height]
- Szerokość trzonu [Stem Width]
- Kolor trzonu [Stem Color]
- Sezon [Season]
- Klasa docelowa [Target Class]

  - Czy jest jadalny, czy nie

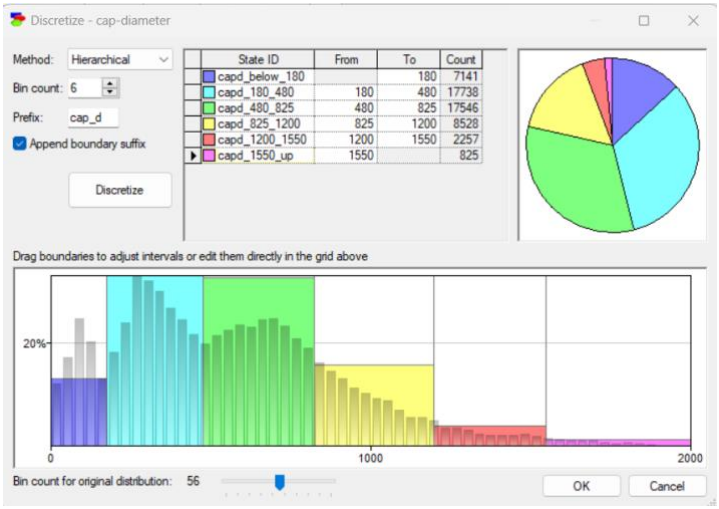
    0 oznacza jadalne

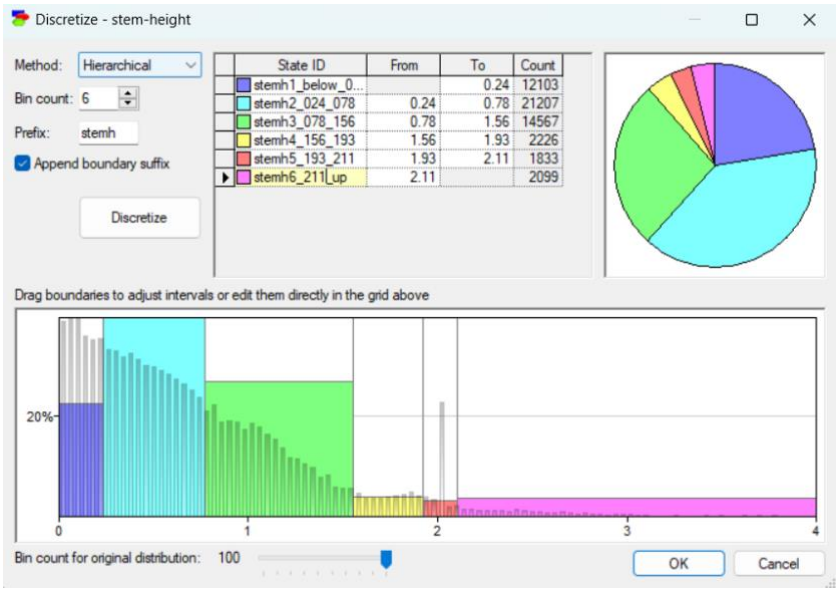
    1 oznacza trujące.

2. Dyskretyzacjaż

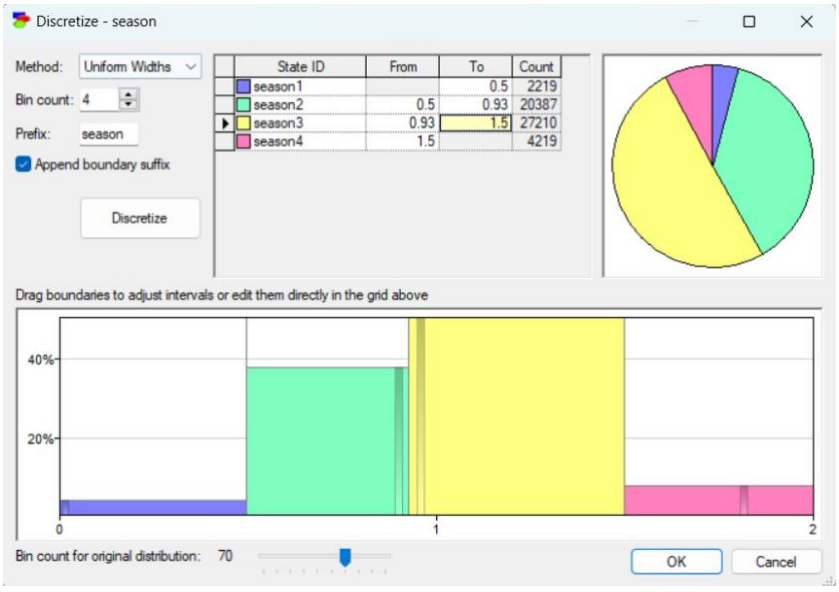
2.1. Opis dyskretyzacji

Dokonując dyskretyzacji największą wagę przykładaliśmy do wyglądu histogramu i próbie dopasowania przedziałów do występujących w nim „górek”, dzięki czemu mniej więcej w jednym przedziale zawierały się grupy rekordów o podobnych specyfikacjach.



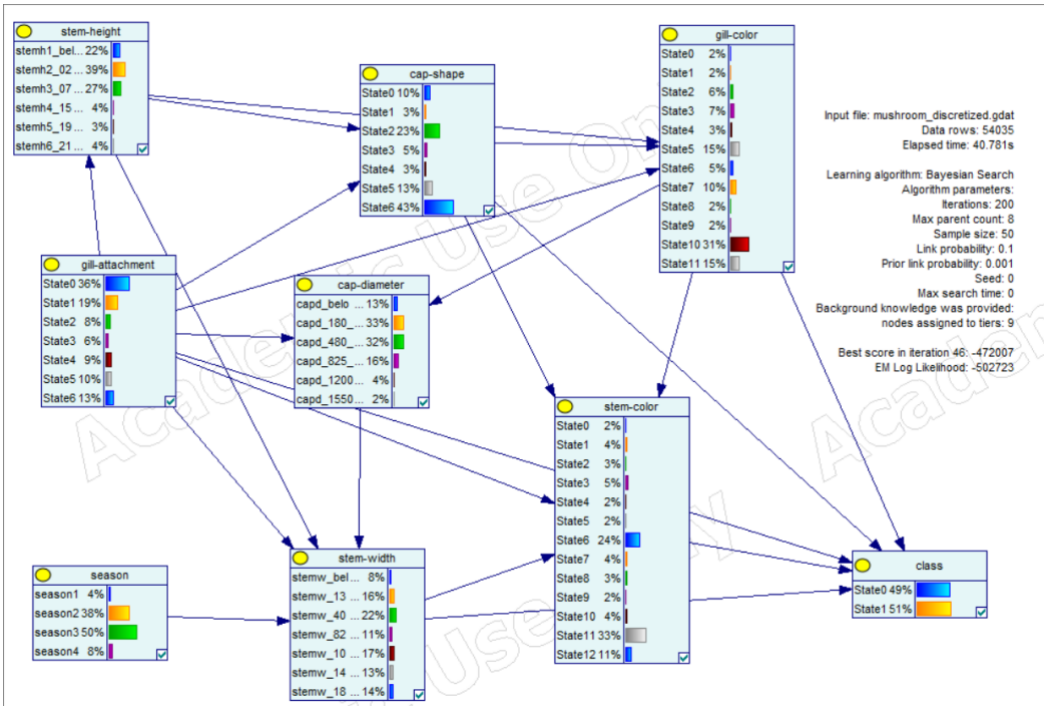

Discretize - stem-height
Method: Hierarchical
Bin count: 6
Prefix: stemh
Append boundary suffix
Discretize
State ID | From | To | Count
stemh1_below_0... | | 0.24 | 12103
stemh2_024_078 | 0.24 | 0.78 | 21207
stemh3_078_156 | 0.78 | 1.56 | 14567
stemh4_156_193 | 1.56 | 1.93 | 2226
stemh5_193_211 | 1.93 | 2.11 | 1833
stemh6_211_up | 2.11 | | 2099
Drag boundaries to adjust intervals or edit them directly in the grid above
20%
0
1
2
3
4
Bin count for original distribution: 100
OK
Cancel


Discretize - stem-width
Method: Hierarchical
Bin count: 7
Prefix: stemw
Append boundary suffix
Discretize
State ID | From | To | Count
stemw_below_135 | | 135 | 4412
stemw_135_400 | 135 | 400 | 8608
stemw_400_825 | 400 | 825 | 12106
stemw_825_1075 | 825 | 1075 | 4928
stemw_1075_1450 | 1075 | 1450 | 9101
stemw_1450_1850 | 1450 | 1850 | 6920
stemw_1850_up | 1850 | | 7960
Drag boundaries to adjust intervals or edit them directly in the grid above
20%
0
1000
2000
3000
4000
Bin count for original distribution: 90
OK
Cancel


Discretize - season
Method: Uniform Widths
Bin count: 4
Prefix: season
Append boundary suffix
Discretize
State ID | From | To | Count
season1 | | 0.5 | 2219
season2 | 0.5 | 0.93 | 20387
season3 | 0.93 | 1.5 | 27210
season4 | 1.5 | | 4219
Drag boundaries to adjust intervals or edit them directly in the grid above
40%
20%
0
1
2
Bin count for original distribution: 70
OK
Cancel

3. Bayesan Search

3.1. Graf



3.2. Statystyki

Basic statistic: Node count:9 Avg indegree: 2.556 Max indegree: 5 Avg outcomes: 7.111 Max outcomes: 13			
Object	**Count**	**States**	**Param./Indep.**
Nodes	9	64	172967 / 113759
Arcs	23	-	-

3.3. Strength of influence

Parent	Child	Average	Maximum	Weighted
gill-attachment	cap-shape	0.386669	0.971219	0.386669
gill-attachment	gill-color	0.36635	0.992707	0.36635
gill-attachment	cap-diameter	0.314784	0.887444	0.314784
gill-attachment	stem-width	0.313075	0.990333	0.313075
cap-diameter	stem-width	0.311458	0.961501	0.311458
cap-shape	gill-color	0.304114	0.890347	0.304114
gill-color	cap-diameter	0.299112	0.909028	0.299112
season	cap-shape	0.295582	0.827397	0.295582
stem-height	cap-shape	0.289089	0.869497	0.289089
stem-height	stem-width	0.244163	0.952365	0.244163
stem-height	gill-color	0.244052	0.955913	0.244052
season	stem-width	0.240748	0.889876	0.240748
gill-attachment	stem-height	0.232832	0.790862	0.232832
season	stem-height	0.182492	0.652162	0.182492

3.4. Diagnosis

3.5. Validation results

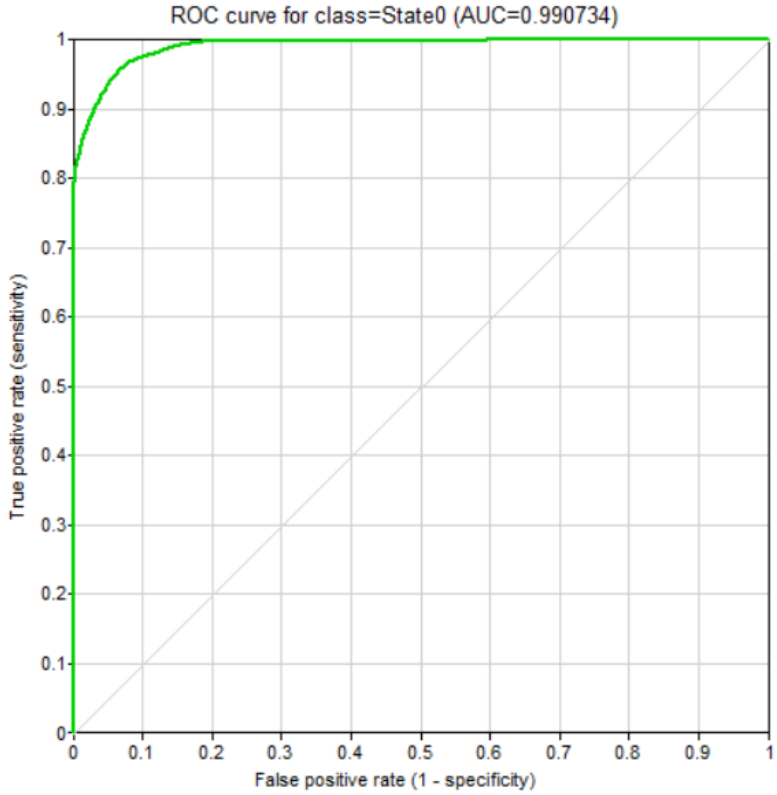
Accuracy:

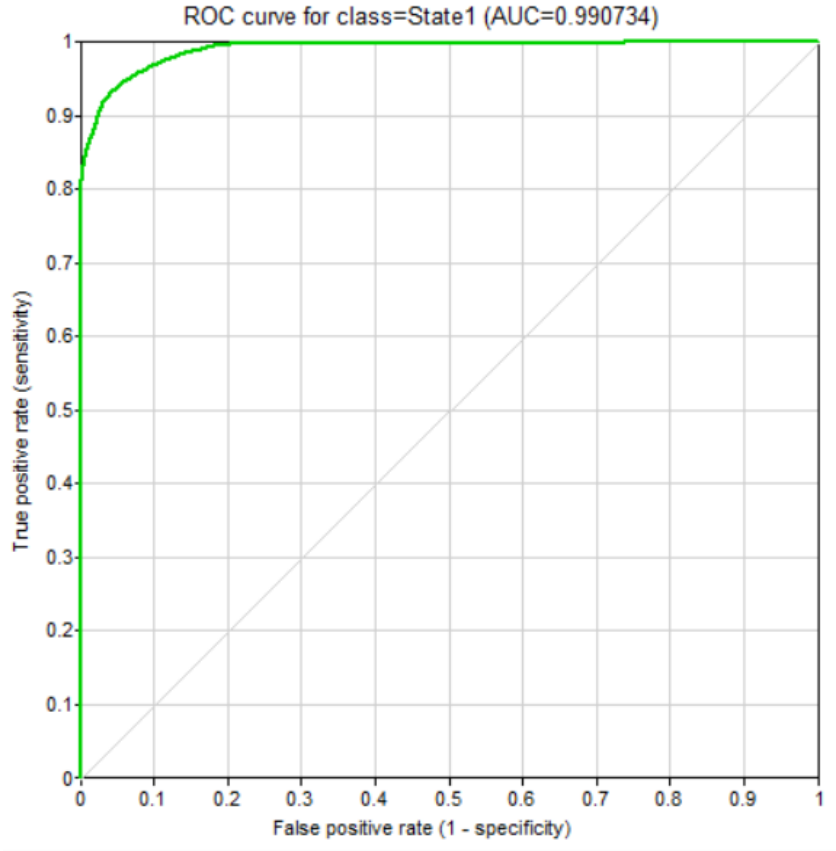
Class = 0.94374 (50995/54035)
State0 = 0.952422 (23201/24360)
State1 = 0.936613 (27794/29675)

3.6. Confusion Matrix

		Predicted	
		State0	State1
Act.	State0	**23201**	1159
	State1	1881	**27794**

3.7. Krzywe ROC

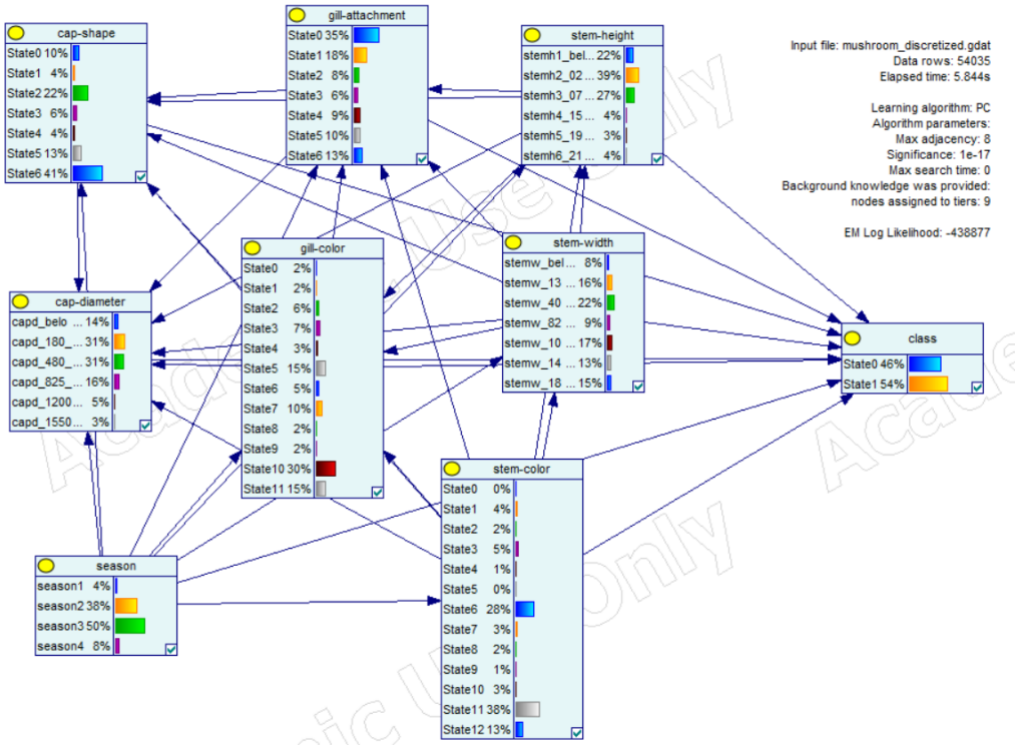




4. PC

Model PC w przypadku wybranego zbioru danych okazał się być mocno problematyczny ze względu na bardzo dużą ilość krawędzi, których nie udało się wyeliminować zmniejszając istotność i dostosowując *background knowledge*. Model jest bardzo duży i plik *genie'go* waży ponad 230Mb. Ze względu na rozmiar modelu program nie wyrabiał i nie było możliwości sprawdzenia mocy wpływu krawędzi oraz przeprowadzić diagnozy modelu (próby podjęto 3 razy i za każdym razem program odmawiał współpracy)

4.1. Graf



4.2. Statystyki

Basic statistic: Node count: 9 Avg indegree: 4 Max indegree: 8 Avg outcomes: 7.111 Max outcomes: 13			
Object	**Count**	**States**	**Param./Indep.**
Nodes	9	64	24611916 / 15410303
Arcs	36	-	-

4.3. Validation results

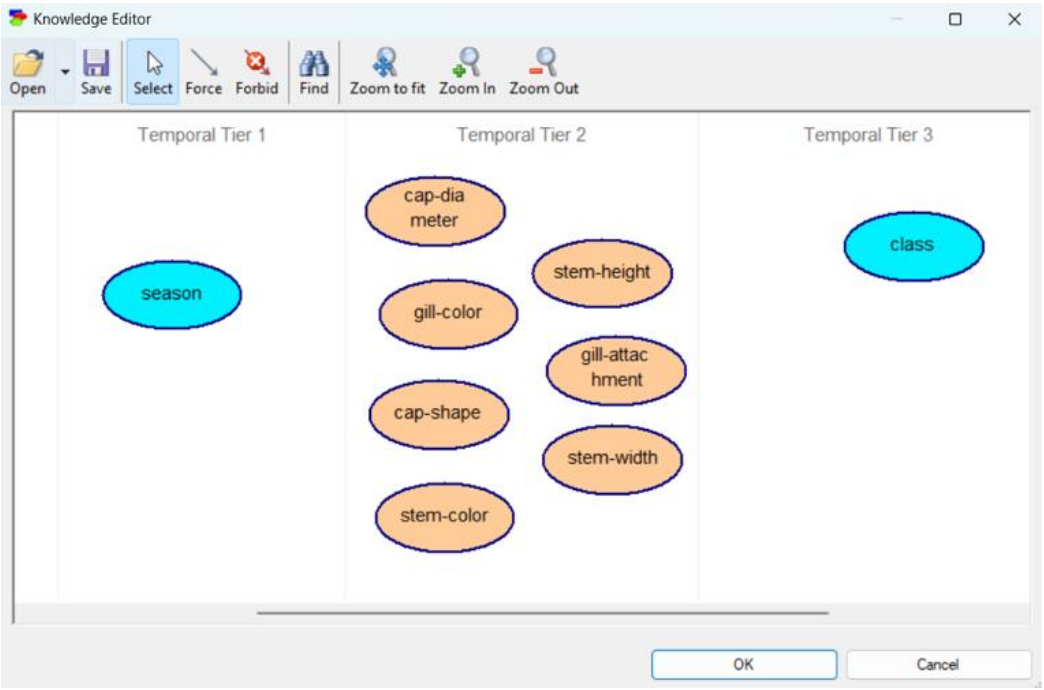
Accuracy:

class = 0.962173 (51991/54035)
State0 = 0.980829 (23893/24360)
State1 = 0.946858 (28098/29675)

4.4. Confusion Matrix

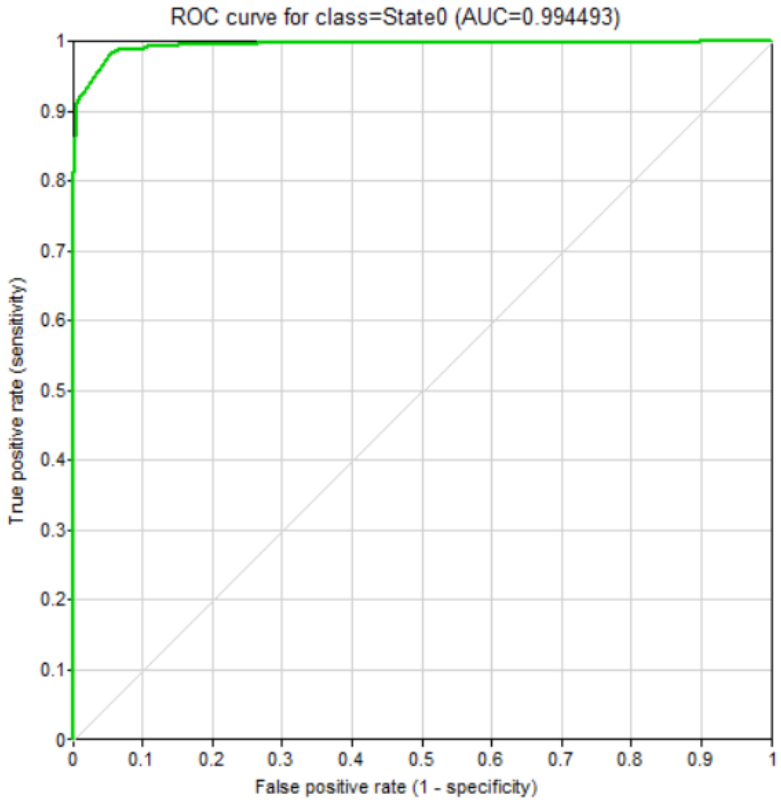
		Predicted	
		State0	State1
Act.	State0	**23893**	467
	State1	1577	**28098**

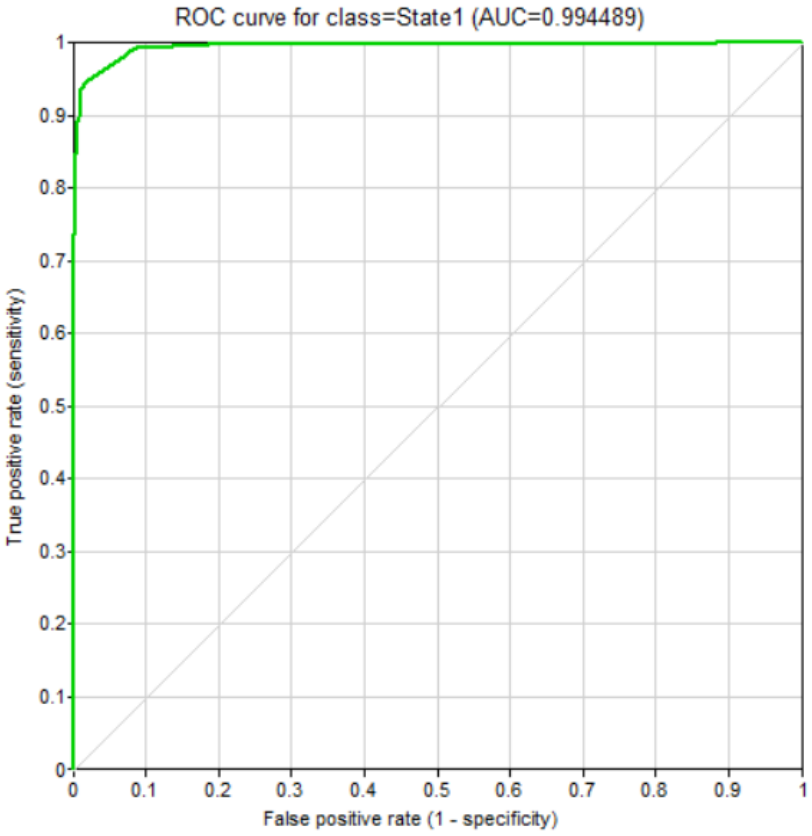
4.5. Background knowledge



W *background knowledge* założono, że żadna cecha grzyba nie ma wpływu na to jaki jest sezon, oraz że na klasę (jadalny, niejadalny) wpływ mogą mieć wszystkie cechy charakterystyczne grzyba.

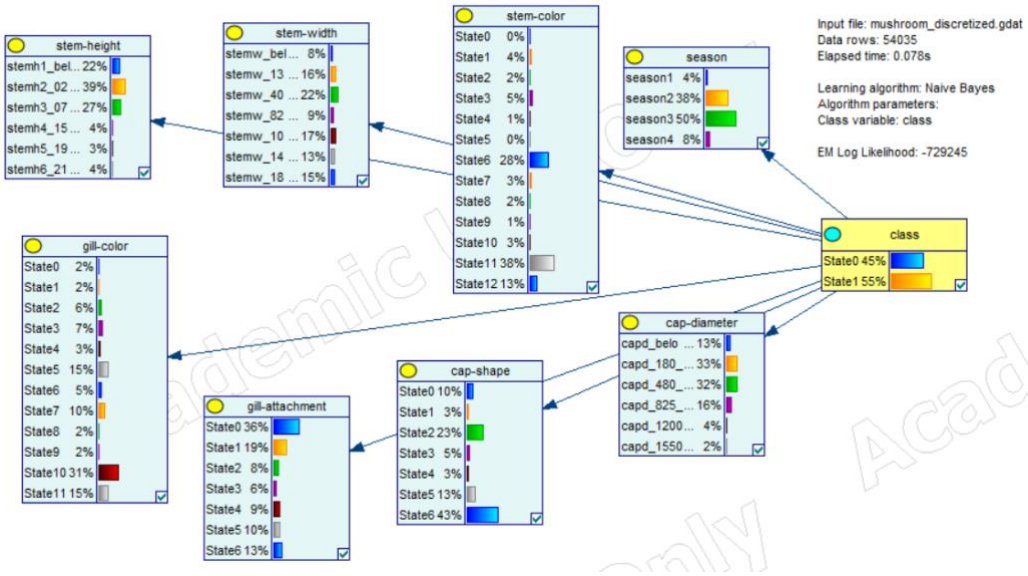
4.6. Krzywe ROC





5. Naive Bayes

5.1. Graf



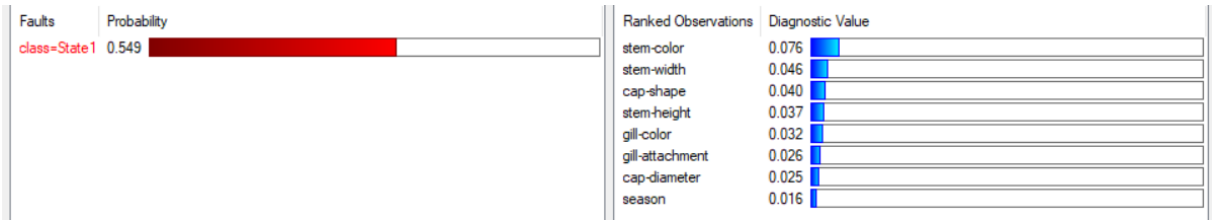
5.2. Statystyki

Basic statistic: Node count: 9 Avg indegree: 0.8889 Max indegree: 1 Avg outcomes: 7.111 Max outcomes: 13			
Object	**Count**	**States**	**Param./Indep.**
Nodes	9	64	126 / 109
Arcs	8	-	-

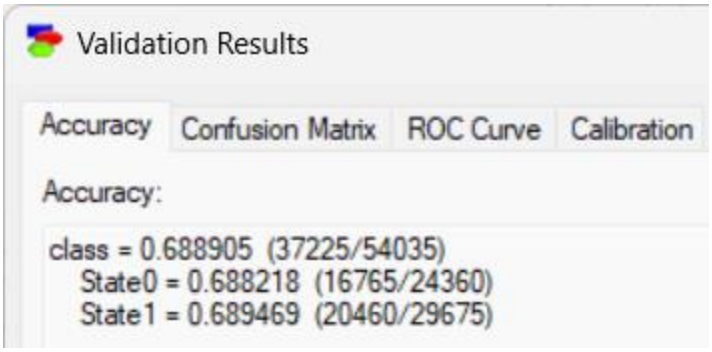
5.3. Strength of influence

P...	Child	Average	Maximum	Weighted
class	stem-color	0.173314	0.173314	0.173314
class	stem-width	0.130041	0.130041	0.130041
class	gill-color	0.120197	0.120197	0.120197
class	stem-height	0.119316	0.119316	0.119316
class	cap-shape	0.113775	0.113775	0.113775
class	cap-diameter	0.112998	0.112998	0.112998
class	gill-attachment	0.0906943	0.0906943	0.0906943
class	season	0.0717646	0.0717646	0.0717646

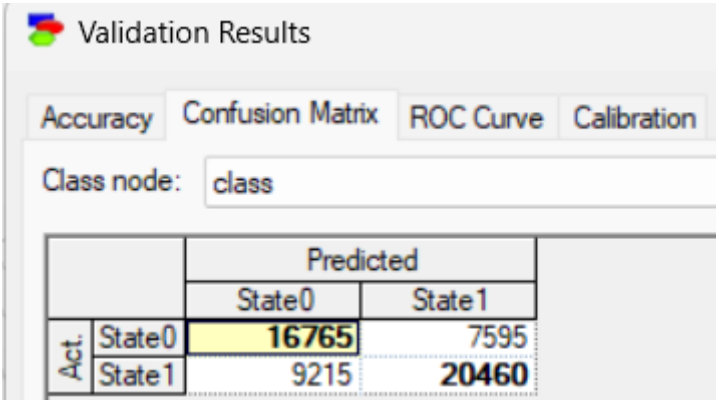
5.4. Diagnosis



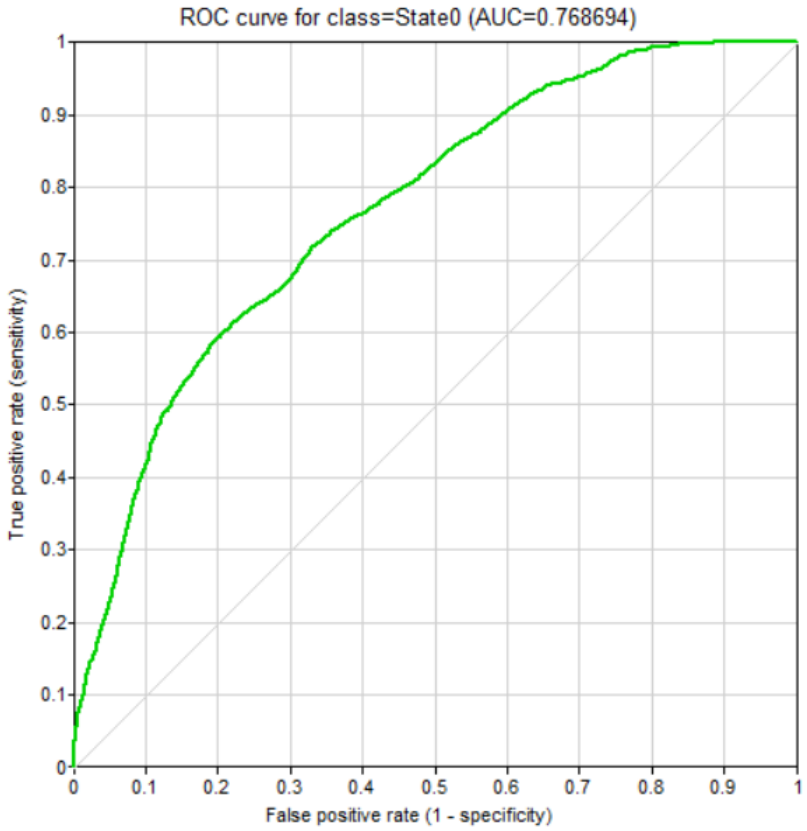
5.5. Validation results

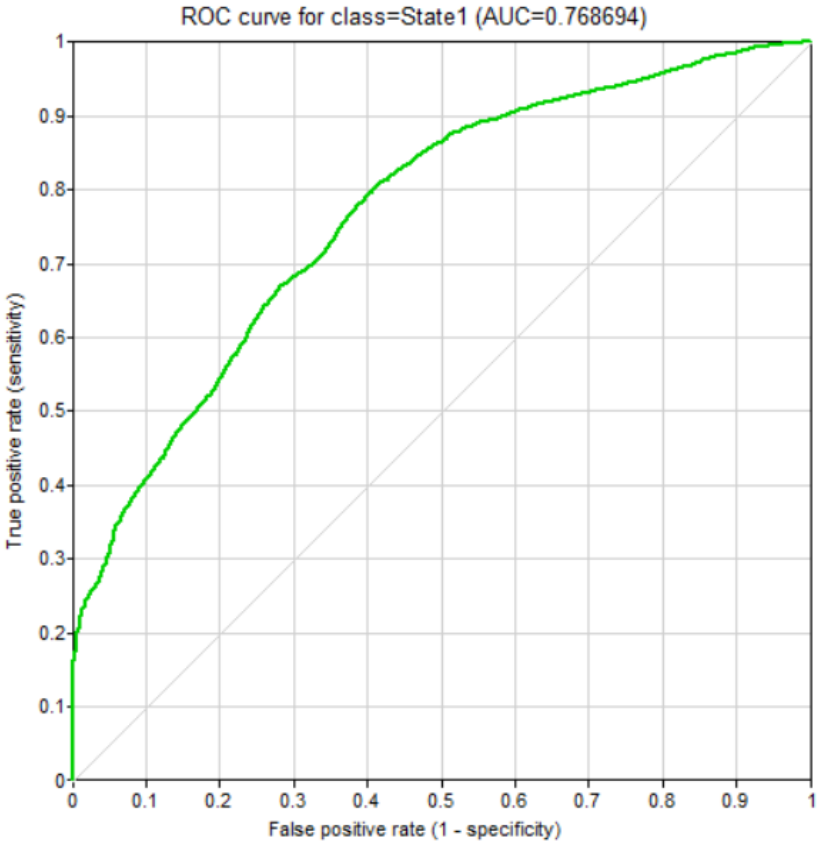


5.6. Confusion Matrix



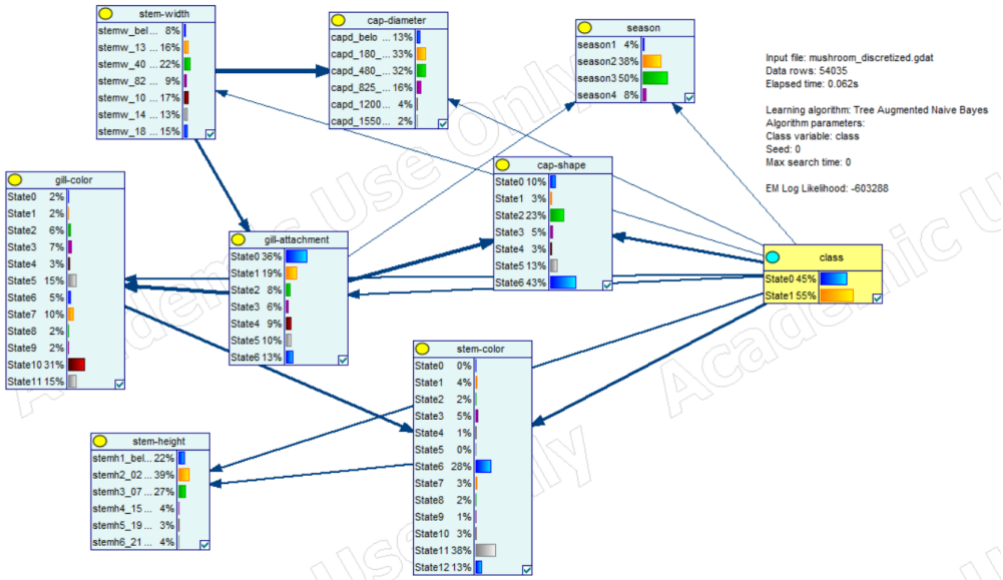
5.7. Krzywe ROC





6. TANB

6.1. Graf



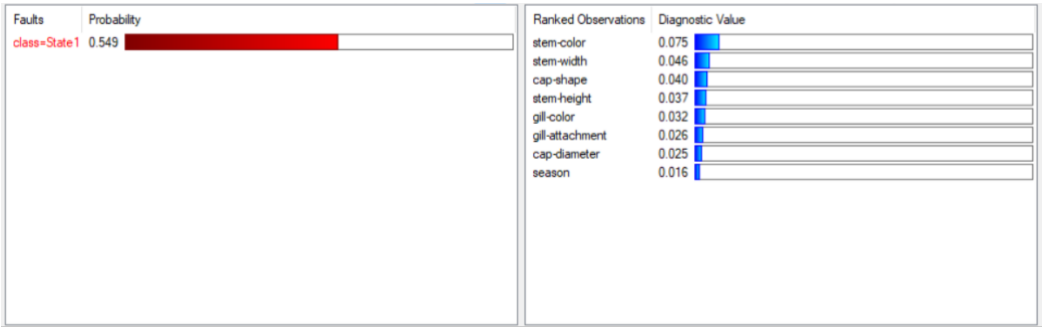
6.2. Statystyki

Basic statistic: Node count: 9 Avg indegree: 1.667 Max indegree: 2 Avg outcomes: 7.111 Max outcomes: 13			
Object	**Count**	**States**	**Param./Indep.**
Nodes	9	64	988 / 865
Arcs	15	-	-

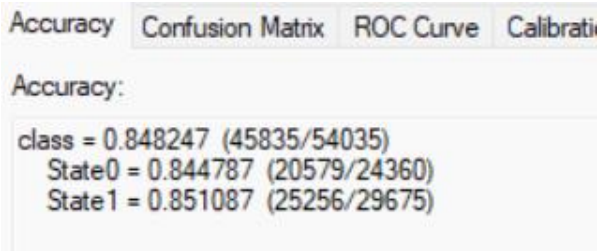
6.3. Strength of influence

Parent	Child	Average	Maximum	Weighted
stem-width	cap-diameter	0.533735	0.903611	0.533735
gill-attachment	gill-color	0.449695	0.874991	0.449695
gill-attachment	cap-shape	0.416207	0.729233	0.416207
gill-color	stem-color	0.389436	0.832181	0.389436
class	stem-color	0.355496	0.661376	0.355496
stem-width	gill-attachment	0.324906	0.677334	0.324906
class	cap-shape	0.319543	0.5699	0.319543
stem-color	stem-height	0.284614	0.973382	0.284614
class	stem-height	0.2585	0.644888	0.2585
class	gill-color	0.238409	0.380963	0.238409
class	gill-attachment	0.211315	0.302929	0.211315
class	cap-diameter	0.167273	0.315082	0.167273
gill-attachment	season	0.153682	0.310825	0.153682
class	season	0.13666	0.23816	0.13666
class	stem-width	0.130041	0.130041	0.130041

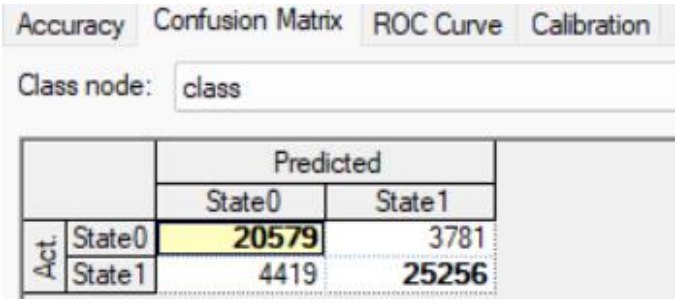
6.4. Diagnosis



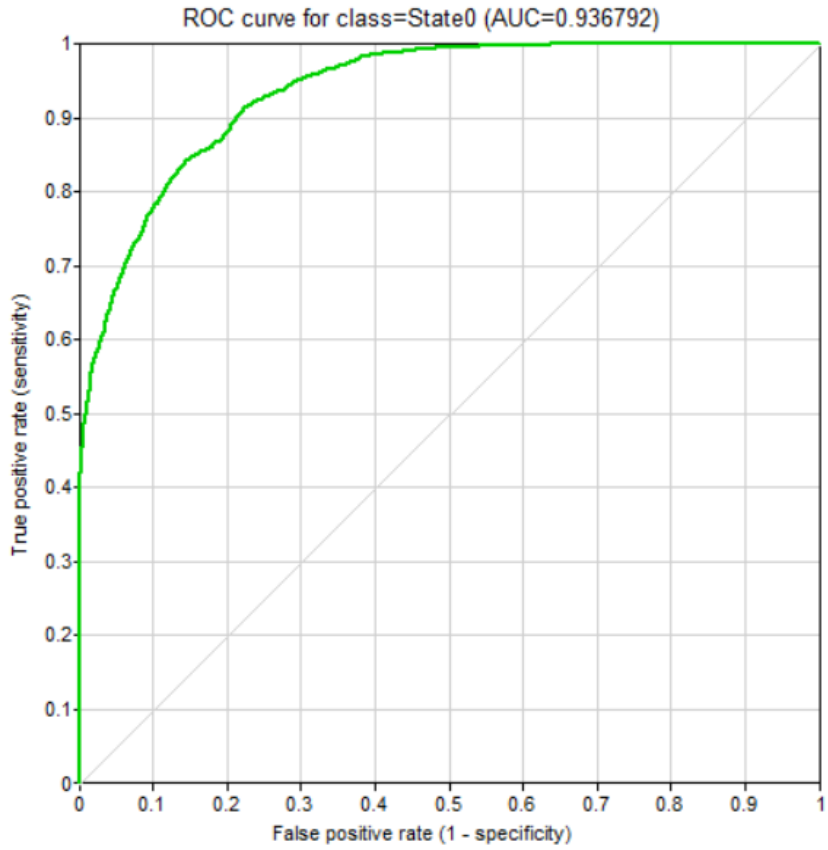
6.5. Validation results

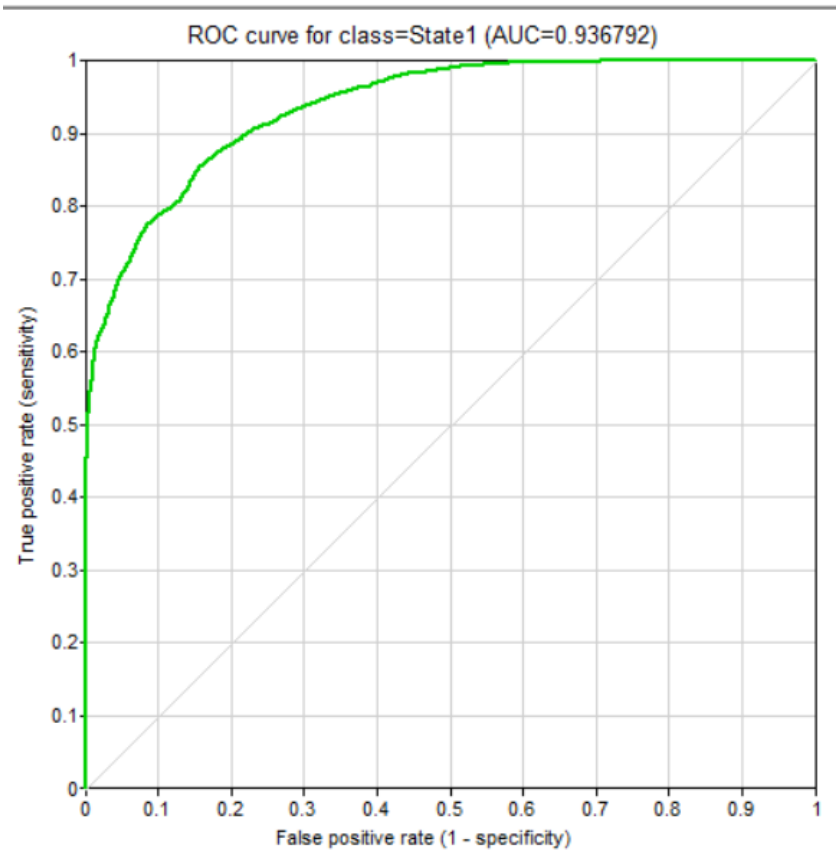


6.6. Confusion Matrix



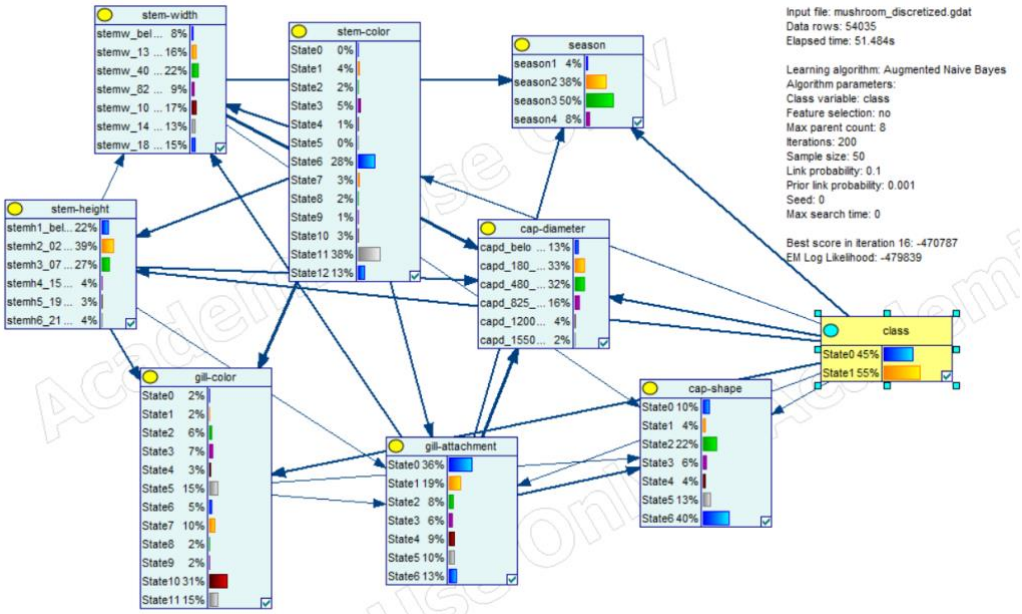
6.7. Krzywe ROC





7. ANB

7.1. Graf



7.2. Statystyki

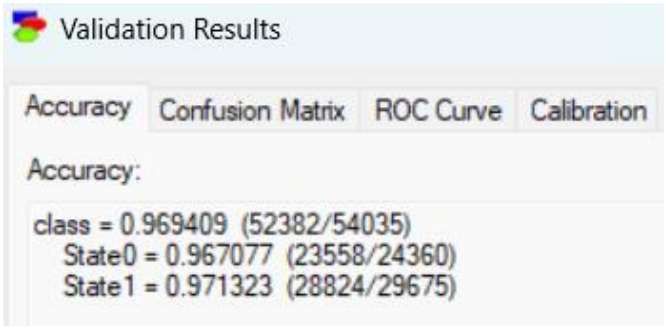
Basic statistic: Node count: 9 Avg indegree: 2.778 Max indegree: 4 Avg outcomes: 7.111 Max outcomes: 13			
Object	**Count**	**States**	**Param./Indep.**
Nodes	9	64	34956 / 29945
Arcs	25	-	-

7.3. Strength of influence

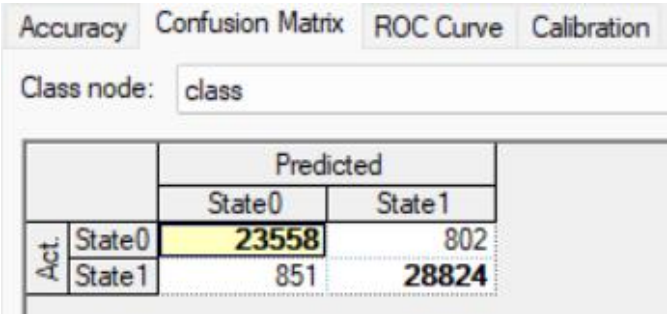
Parent	Child	Average	Maximum	Weighted
gill-attachment	cap-shape	0.386669	0.971219	0.386669
gill-attachment	gill-color	0.36635	0.992707	0.36635
gill-attachment	cap-diameter	0.314784	0.887444	0.314784
gill-attachment	stem-width	0.313075	0.990333	0.313075
cap-diameter	stem-width	0.311458	0.961501	0.311458
cap-shape	gill-color	0.304114	0.890347	0.304114
gill-color	cap-diameter	0.299112	0.909028	0.299112
season	cap-shape	0.295582	0.827397	0.295582
stem-height	cap-shape	0.289089	0.869497	0.289089
stem-height	stem-width	0.244163	0.952365	0.244163
stem-height	gill-color	0.244052	0.955913	0.244052
season	stem-width	0.240748	0.889876	0.240748
gill-attachment	stem-height	0.232832	0.790862	0.232832
season	stem-height	0.182492	0.652162	0.182492

7.4. Diagnosis

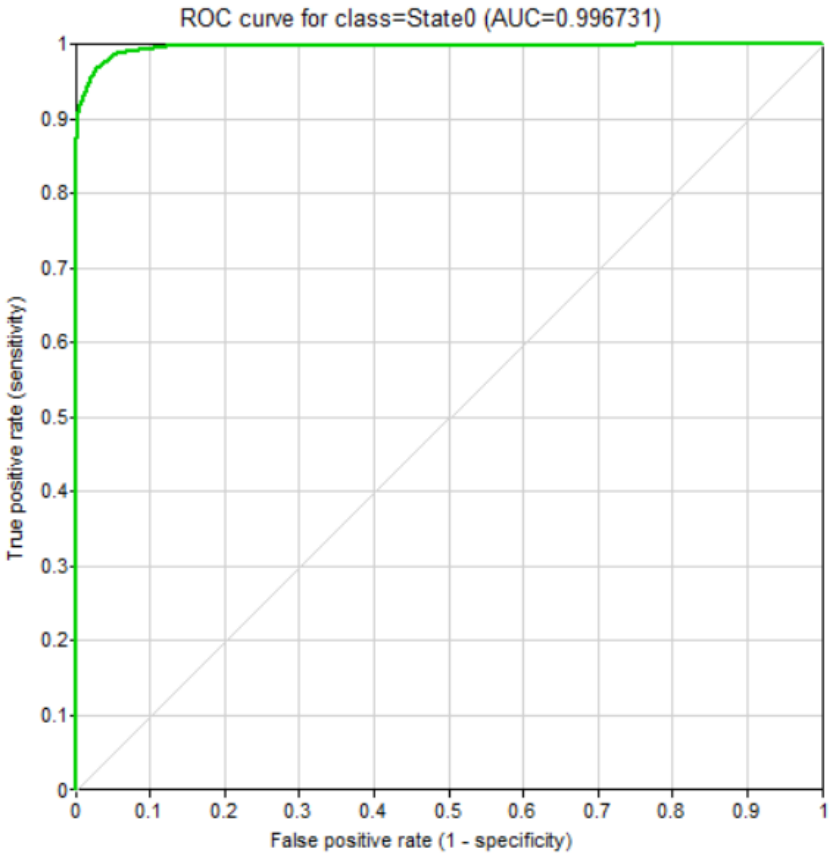
7.5. Validation results

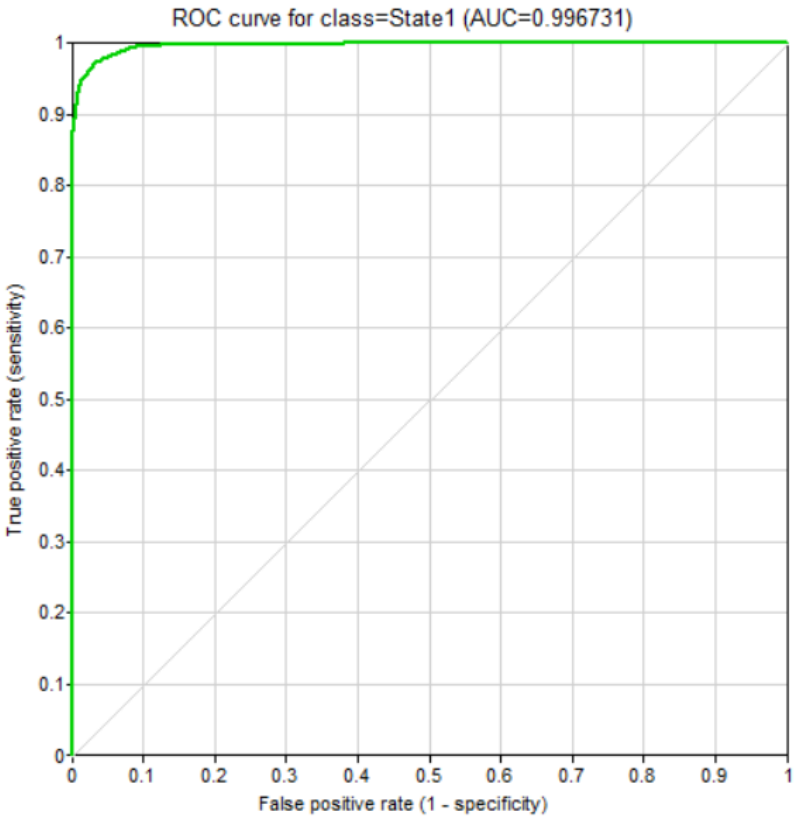


7.6. Confusion Matrix



7.7. Krzywe ROC





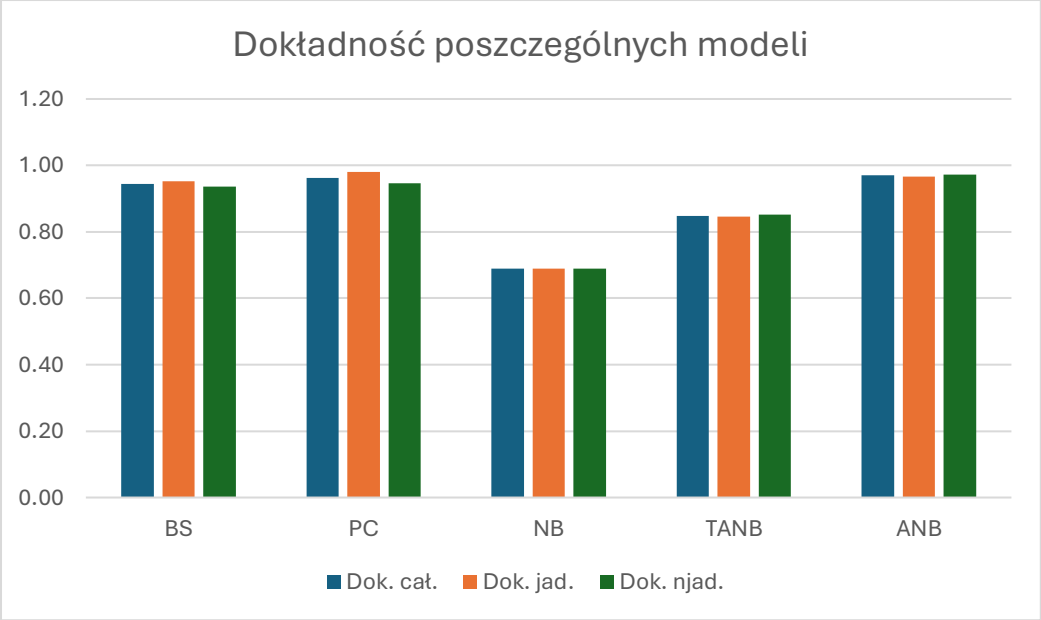
8. Porównanie wyników konkretnych modeli

Modele zostały porównane na podstawie wyników osiągniętych podczas walidacji metodą krzyżową (k-fold, gdzie **k = 10**).

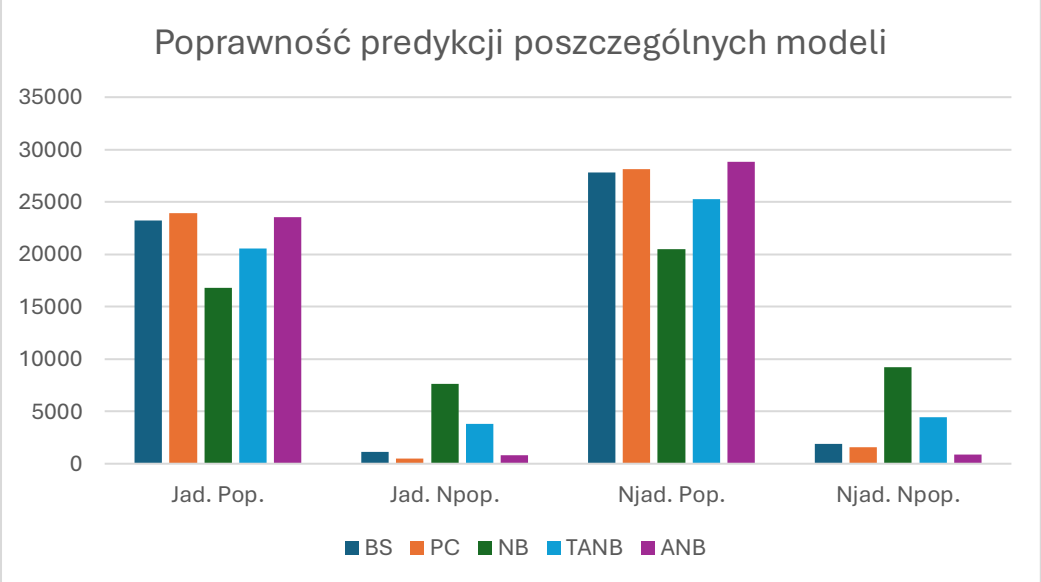
8.1. Tabela porównująca modele

Sieć	Dok. cał.	Dok. jad.	Dok. njad.	Jad. Pop.	Jad. Npop.	Njad. Pop.	Njad. Npop.
BS	0,94	0,95	0,94	23201	1159	27794	1881
PC	0,96	0,98	0,95	23893	467	28098	1577
NB	0,69	0,69	0,69	16765	7595	20460	9215
TANB	0,85	0,84	0,85	20579	3781	25256	4419
ANB	0,97	0,97	0,97	23558	802	28824	851

8.2. Porównanie dokładności



8.3. Porównanie predykcji



8.4. Wnioski i wyniki porównania

Nie wszystkie modele klasyfikacyjne nadają się do każdego zbioru danych. Jak wynika z analizy, modele takie jak Naive Bayes (NB) i Tree Augmented Naive Bayes (TANB) nie osiągnęły zadowalającej dokładności w klasyfikacji grzybów, w przeciwieństwie do Augmented Naive Bayes (ANB) i PC, które uzyskały znacznie lepsze wyniki.

Aby poprawnie wybrać i dostosować model do zbioru danych, konieczne jest głębokie zrozumienie dziedziny, której dane dotyczą. W przypadku klasyfikacji grzybów, wiedza na temat biologicznych cech grzybów i ich wzajemnych relacji może pomóc w lepszym modelowaniu i wyborze cech istotnych dla klasyfikacji. Przykładem może być sezonowość grzybów, która może wpływać na ich cechy, co jest uwzględnione w background knowledge dla modelu PC.

Model Naive Bayes z założeniem niezależności cech nie jest odpowiedni dla zbioru danych, w którym cechy mogą być silnie skorelowane. Wynika to z faktu, że cechy grzybów mogą mieć zależności, które NB nie jest w stanie uchwycić. Patrząc na wykresy porównujące modele na podstawie ich dokładności i predykcji można zauważyć, że modele **NB** oraz **TANB** znacząco odstają wynikami od reszty. Ich dokładność nie przekracza *90%*, a w przypadku **NB** nawet *70%*. Pozostałe sieci poradziły sobie całkiem dobrze osiągając dokładności powyżej *94%*.

Najwyższą dokładność **całkowitą** osiągnął model oparty o algorytm **ANB** – ***97%***
Najwyższą dokładność **jadalnych** osiągnął model oparty o algorytm **PC** – ***98%***
Najwyższą dokładność **niejadalnych** osiągnął model oparty o algorytm **ANB** – ***97%***