Politechnika Białostocka

Sprawozdanie z pracowni specjalistycznej

Zaawansowane bazy danych i

hurtownie danych

Pracownia specjalistyczna numer: 13-14

Temat: Projekt- Analiza Danych w oparciu o modele sieci Bayesowskich

Wykonujący ćwiczenie:

• Michał Wołosewicz

• Patryk Wójtowicz

Studia dzienne

Kierunek: Informatyka II stopień

Semestr: I

Grupa zajęciowa: PS 4

Prowadzący ćwiczenie: dr hab. inż. prof. PB Agnieszka Drużdżel

Data wykonania ćwiczenia: 16.06.2024

1. Krótki opis analizowanego zbioru danych

Zbiór danych pochodzi z rocznej ankiety BRFSS prowadzonej przez CDCP (Centers for Disease Control and Prevention) w 2015 roku. Ankieta zbiera odpowiedzi od ponad 400,000 Amerykanów (w projekcie wykorzystano pierwszych 10,119 ankietowanych) i dotyczy ona zachowań ryzykownych dla zdrowia, przewlekłych chorób, korzystania z usług profilaktyki zdrowotnej oraz parametrów ankietowanej jednostki. Analiza ma na celu wykrycie przyczyn występowania cukrzycy i jakie wskaźniki są za nią odpowiedzialne.

Wskaźniki wykorzystane w zbiorze danych:

1. Diabetes_012 – opisuje stan cukrzycy u aniektowanej osoby:
    - 0 – brak cukrzycy
    - 1 – stan przed cukrzycowy
    - 2 – cukrzyca

2. HighBP – określa czy ankietowana osoba posiada wysokie ciśnienie krwi:
    - 0 – nie
    - 1 – tak

3. HighChol – określa czy ankietowana osoba posiada wysoki cholesterol:
    - 0 – nie
    - 1 – tak

4. CholCheck – czy badano cholesterol chociaż raz w ciągu ostatnich 5 lat:
    - 0 – nie
    - 1 – tak

5. BMI – wskaźnik BMI masy ciała.

6. Smoker – czy ankietowana osoba zapaliła więcej niż 100 papierosów w ciągu swojego całego życia:
    - 0 – nie
    - 1 – tak

7. Stroke – określa czy ankietowana osoba przeszła udar:
    - 0 – nie
    - 1 – tak

8. HeartDiseaseorAttack – określa czy u ankietowanej osoby występowała choroba niedokrwienna serca lub zawał mięśna sercowego:
    - 0 – nie
    - 1 – tak

9. PhysActivity – określa czy ankietowana osoba posiadała jakąkolwiek aktywność fizyczną nie związaną z wykonywanym zawodem przez ostatnie 30 dni:
    - 0 – nie
    - 1 – tak

10. Fruits – czy jednostka spożywa jedno lub więcej owoców dziennie:
    - 0 – nie
    - 1 – tak

11. Veggies – czy jednostka spożywa jedno lub więcej warzyw dziennie:
    - 0 – nie
    - 1 – tak

12. HeavyAlcoholConsmup – określa czy mężczyzna spożywa więcej niż 14 napoi alkoholowych w tygodniu a kobieta więcej niż 7 napoi alkoholowych w tygodniu:
    - 0 – nie
    - 1 – tak

13. AnyHealthcare – określa posiadanie jakiejkolwiek ochrony zdrowia jak ubezpieczenie zdrowotne przez jednostkę:
    - 0 – nie
    - 1 – tak

14. NoDocbcCost – ankietowany zrezygnował z wizyty u lekarza i leczenia z powodu zbyt wysokich koszótw finansowych jakie się z tym wiązały:
    - 0 – nie
    - 1 – tak

15. GenHlth – ankietowany ocenia swój stan zdrowia:
    - 1 – wspaniały
    - 2 – bardzo dobry
    - 3 – dobry
    - 4 – zadowalający
    - 5 – słaby

16. MentHlth – ilość dni w których odnotowano problemy zdrowia psychicznego z ostatnich 30 dni. Wynik to liczba z zakresu od 0 do 30.

17. PhysHlth – ilość dni w których odnotowano problemy, choroby lub urazy zdrowia fizycznego z ostatnich 30 dni. Wynik to liczba z zakresu od 0 do 30.

18. DiffWalk – określa czy ankietowany ma trudności z poruszaniem się po schodach i terenach płaskich:
    - 0 - nie
    - 1 - tak

19. Sex – płeć:
    - 0 - kobieta
    - 1 - mężczyzna

20. Age – grupa wiekowa:
    - 1 odpowiada za wiek od 18 do 24 lat
    - 2 odpowiada za wiek od 25 do 29 lat

- 3 odpowiada za wiek od 30 do 34 lat
- 4 odpowiada za wiek od 35 do 39 lat
- 5 odpowiada za wiek od 40 do 44 lat
- 6 odpowiada za wiek od 45 do 49 lat
- 7 odpowiada za wiek od 50 do 54 lat
- 8 odpowiada za wiek od 55 do 59 lat
- 9 odpowiada za wiek od 60 do 64 lat
- 10 odpowiada za wiek od 65 do 69 lat
- 11 odpowiada za wiek od 70 do 74 lat
- 12 odpowiada za wiek od 75 do 79 lat
- 13 odpowiada za wiek od 80 w górę

21. Education – stopień wykształcenia:
    - 1 - przedszkole
    - 2 - podstawówka
    - 3 - gimnazjum
    - 4 - szkoła średnia
    - 5 - pierwszy stopień studiów
    - 6 - drugi stopeiń studiów lub wyżej

22. Income – zarobki:
    - 1 odpowiada za przedział od 0$ do 9,999$
    - 2 odpowiada za przedział od 10,000$ do 14,999$
    - 3 odpowiada za przedział od 15,000$ do 19,999$
    - 4 odpowiada za przedział od 20,000$ do 24,999$
    - 5 odpowiada za przedział od 25,000$ do 34,999$
    - 6 odpowiada za przedział od 35,000$ do 49,999$
    - 7 odpowiada za przedział od 50,000$ do 74,999$
    - 8 odpowiada za przedział od 75,00$ w górę

2. Szczegóły dyskretyzacji zmiennych ciągłych

- BMI – wskaźnik został podzielony na 8 grup:
    1. Poniżej 16.00 – wygłodzenie
    2. Od 16.00 do 16.99 – wychudzenie
    3. Od 17.00 do 18.49 – niedowaga
    4. Od 18.50 do 24.99 – wartość prawidłowa
    5. Od 25.00 do 29.99 – nadwaga
    6. Od 30.00 do 34,99 – otyłość I stopnia
    7. Od 35.00 do 39,99 – otyłość II stopnia
    8. Powyżej 40.00 – otyłość III stopnia

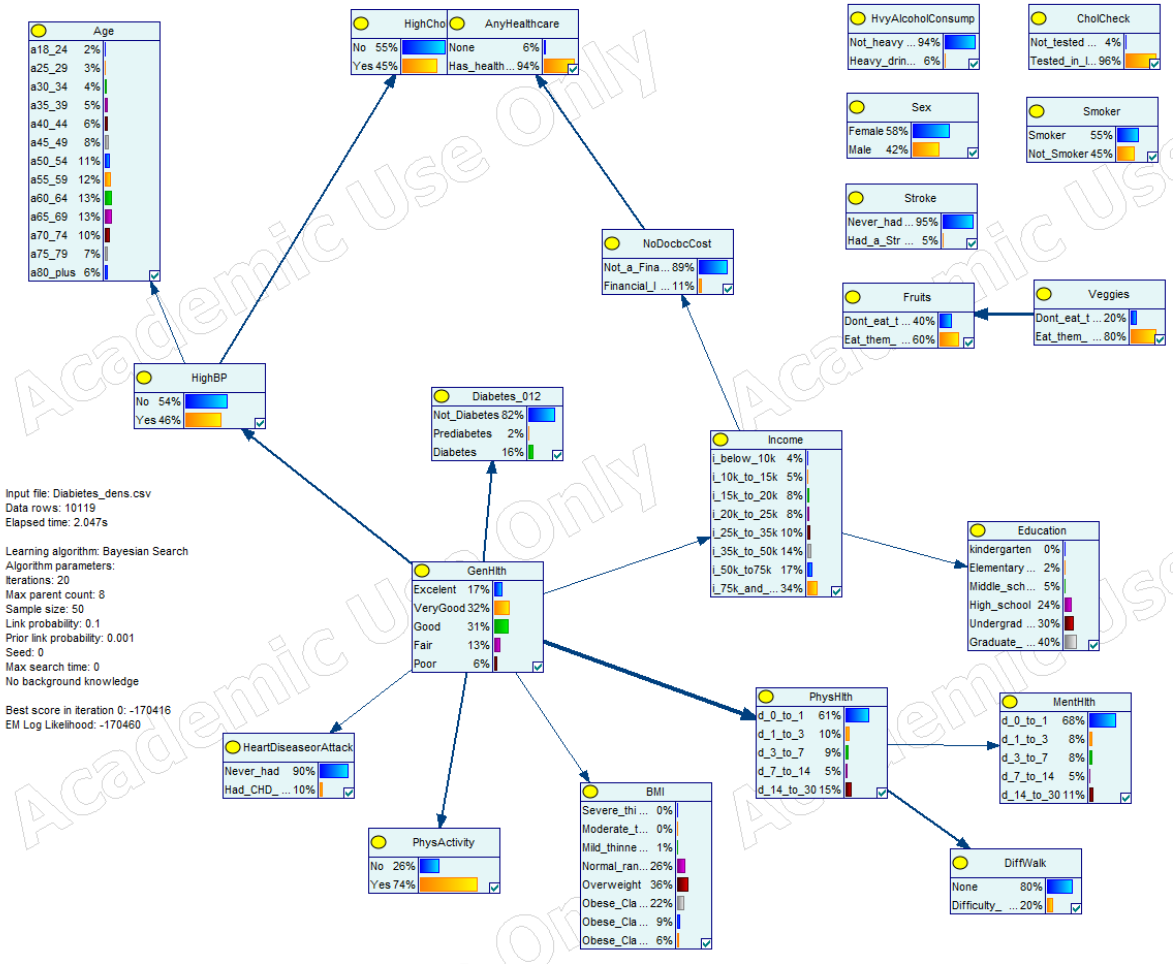
- MentHlth – wskaźnik został podzielony na 5 grup:
    1. d_0_to_1 – dla 0 dni
    2. d_1_to_3 – od 1 do 3 dni
    3. d_3_to_7 – od 4 do 7 dni

4. d_7_to_14 – od 7 do 14 dni
5. d_14_to_30 – od 15 do 30 dni

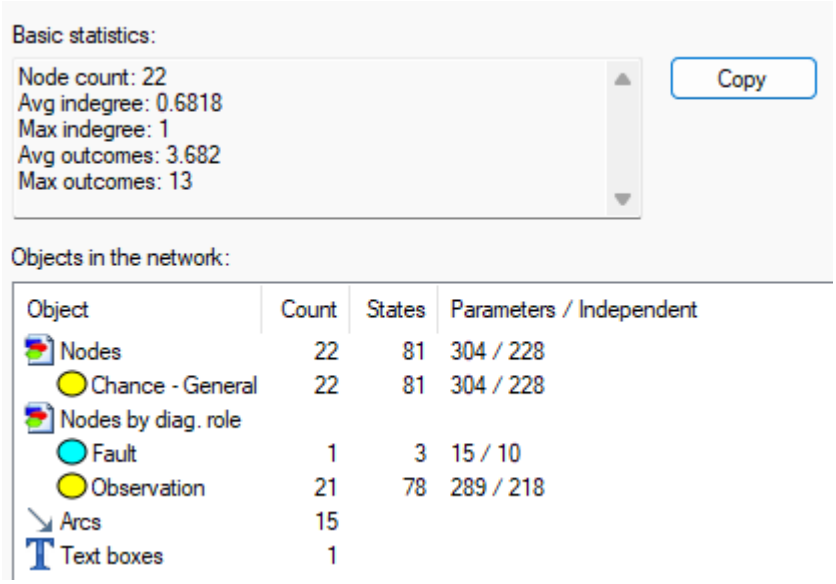
- PhysHlth – wskaźnik został podzielony na 5 grup:
  1. d_0_to_1 – dla 0 dni
  2. d_1_to_3 – od 1 do 3 dni
  3. d_3_to_7 – od 4 do 7 dni
  4. d_7_to_14 – od 7 do 14 dni
  5. d_14_to_30 – od 15 do 30 dni

3. Modele

- Bayesian Search:



*Rysunek 1 Struktura przyczynowo- skutkowa **Bayesian Search** z węzłami w wersji prezentacji z rozkładami brzegowymi.*



Rysunek 2 Podstawowe dane o modelu Bayesian Search.

Liczba węzłów: 22
Średnia liczba rodziców węzła (średnia liczba stopni): 0,6818
Maksymalna liczba rodziców węzła (maks. Liczba stopni): 1
Średnia liczba wyników węzłów: 3,682
Maksymalna liczba wyników węzłów: 13
Ilość krawędzi: 15

Strength of Influence

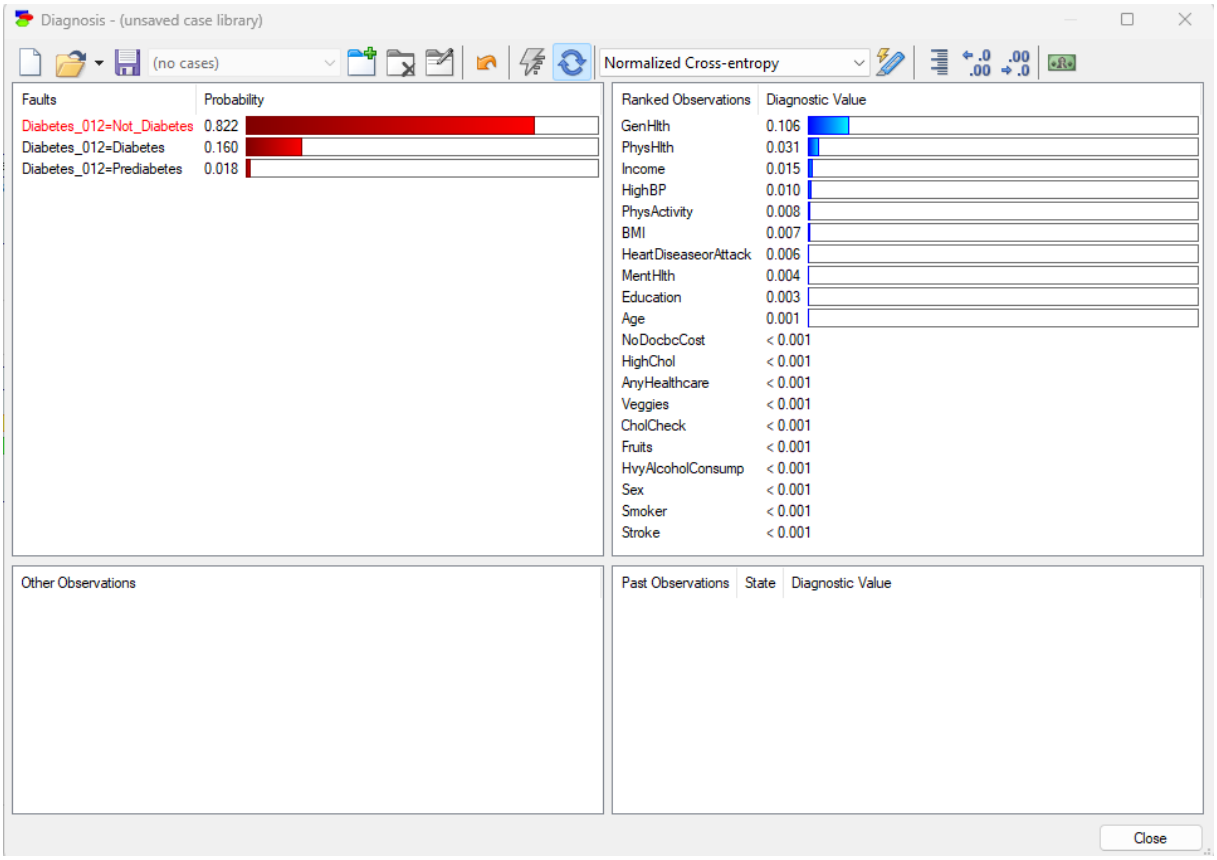
Arcs:

Parent	Child	Average	Maximum	Weighted
GenHlth	PhysHlth	0.421127	0.801435	0.421127
Veggies	Fruits	0.326121	0.326121	0.326121
HighBP	HighChol	0.286613	0.286613	0.286613
GenHlth	HighBP	0.265267	0.511042	0.265267
PhysHlth	DiffWalk	0.260904	0.545946	0.260904
GenHlth	PhysActivity	0.237168	0.474887	0.237168
NoDocbcCost	AnyHealthcare	0.207119	0.207119	0.207119
GenHlth	Diabetes_012	0.20097	0.375758	0.20097
Income	Education	0.19613	0.436725	0.19613
GenHlth	Income	0.195122	0.355719	0.195122
PhysHlth	MentHlth	0.184636	0.33137	0.184636
GenHlth	HeartDiseaseorAttack	0.15648	0.320199	0.15648
GenHlth	BMI	0.123	0.213304	0.123
Income	NoDocbcCost	0.119155	0.282091	0.119155
HighBP	Age	0.112434	0.112434	0.112434

*Rysunek 3 Lista krawędzi modelu **Bayesian Search** z najsilniejszą mocą wpływu.*

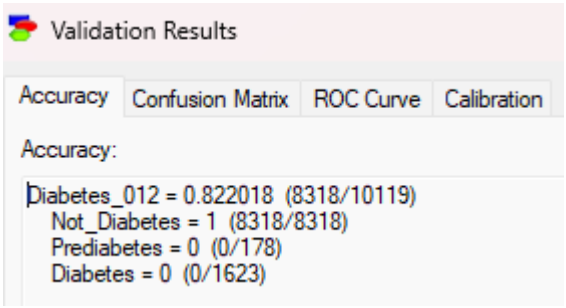
W modelu Bayesian Search krawędź z najsilniejszą mocą wpływu to krawędź **GenHealth - PhysHlth** (generalna opinia o zdrowiu i problemy ze zdoriwem fizycznym w ciagu ostatnich 30 dni). Kolejne krawędzie z bardzo dużą mocą wpływu to **Veggies - Fruits**

(spozywanie cddzienne warzyw oraz spożywanie cddzienne owoców), **HighBP - HighChol** (wysokie ciśnienie krwi oraz wysoki cholesterol), **GenHlth - HighBP** (ocena samodzielna stanu zdrowia i wysokie ciśnienie krwi). Głównie od tych połączeń zależy przewidywanie czy osoba może mieć cukrzyce, czy jest na to duże prawdopodobieństwo. Reszta połączeń ma tutaj mniejsze znaczenie od wcześniej wymienionych jednakże posiadają wysokie współczynniki. Najmniej znaczącym połączeniem jest krawędź **HighBP – Age** (wysokie ciśnienie krwi oraz grupa wiekowa).



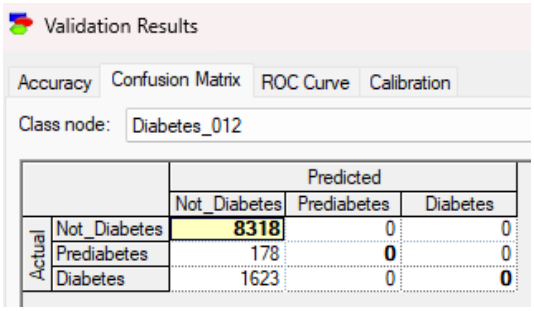
*Rysunek 4 Lista zmiennych modelu **Bayesian Search**, które mają najwyższą wartość diagnostyczną.*

Powyżej znajduje się lista zmiennych modelu **Bayesian Search**, które mają najwyższą wartość diagnostyczną w celu różnicowania zmiennej, która reprezentuje klasę Diabetes_012 (cukrzyca). Najwyższą wartość diagnostyczną uzyskała zmienna *GenHlth* (ocena samodzielna stanu zdrowia) uzyskując wynik 0.106. Kolejną zmienną o wysokiej wartości diagnostycznej jest: *PhysHlth* (urazy zdrowia fizycznego z ostatnich 30 dni) z wynikiem 0.031. Reszta zmiennych ma mniejszą wartość diagnostyczną od wymienionych wyżej zmiennych.



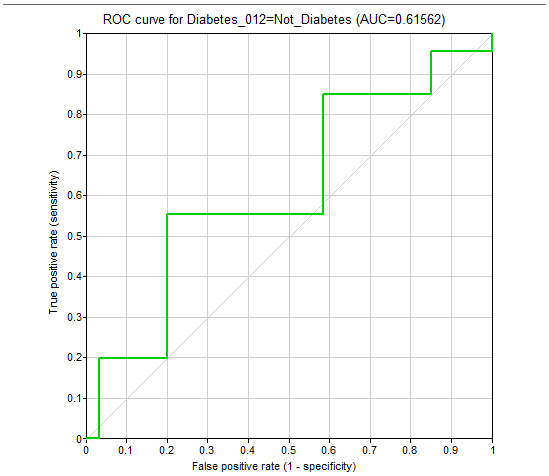
*Rysunek 5 Lista Dokładności dla modelu **Bayesian Search**.*

Jakość modelu **Bayesian Search** wynosi ~0.822 czyli około 82,2%. Został on wyliczony na podstawie ilości poprawnie przewidzianych odpowiedzi czy osoba ma cukrzyce, czy jest w tanie przecukrzycowym lub nie ma cukrzycy. Model odgadł 8318 na 10119 przypadków. Jest to wynik najlepszy wśród badanych modeli. Przy sprawdzaniu jakości modelu zastosowano metodę **Leave one out**.

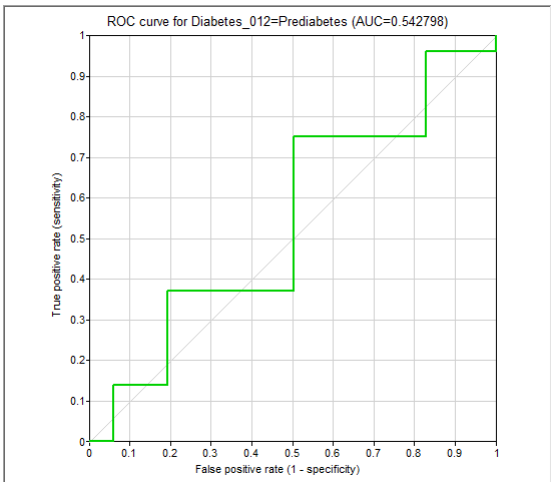


*Rysunek 6 Macierz Pomyłek dla modelu **Bayesian Search**.*

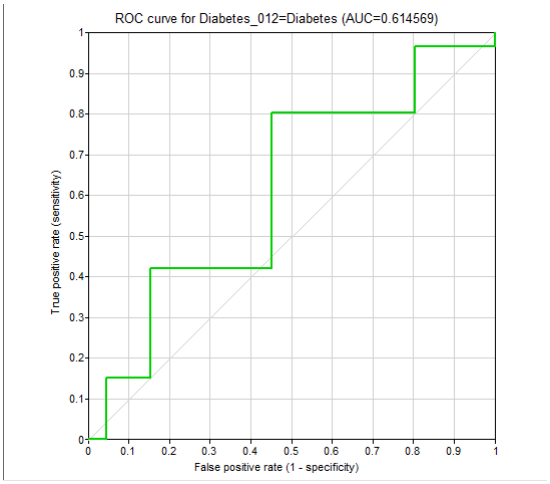
Z macierzy pomyłek dla modelu **Bayesian Search** wynika, że dla wartości “Not_Diabetes” zostało poprawnie przewidzianych 8318 odpowiedzi, a błędnie 0. Dla odpowiedzi “Prediabetes” zostało poprawnie przewidzianych 0 odpowiedzi, a błędnie 178. Dla odpowiedzi “Diabetes” zostało poprawnie przewidzianych 0 odpowiedzi, a błędnie 1623. Łącznie model przewidział poprawnie 8318 wartości, a błędnie 1801 odpowiedzi.



*Rysunek 7 Krzywa ROC dla modelu **Bayesian Search** dla odpowiedzi **Nie_Cukrzyk**.*

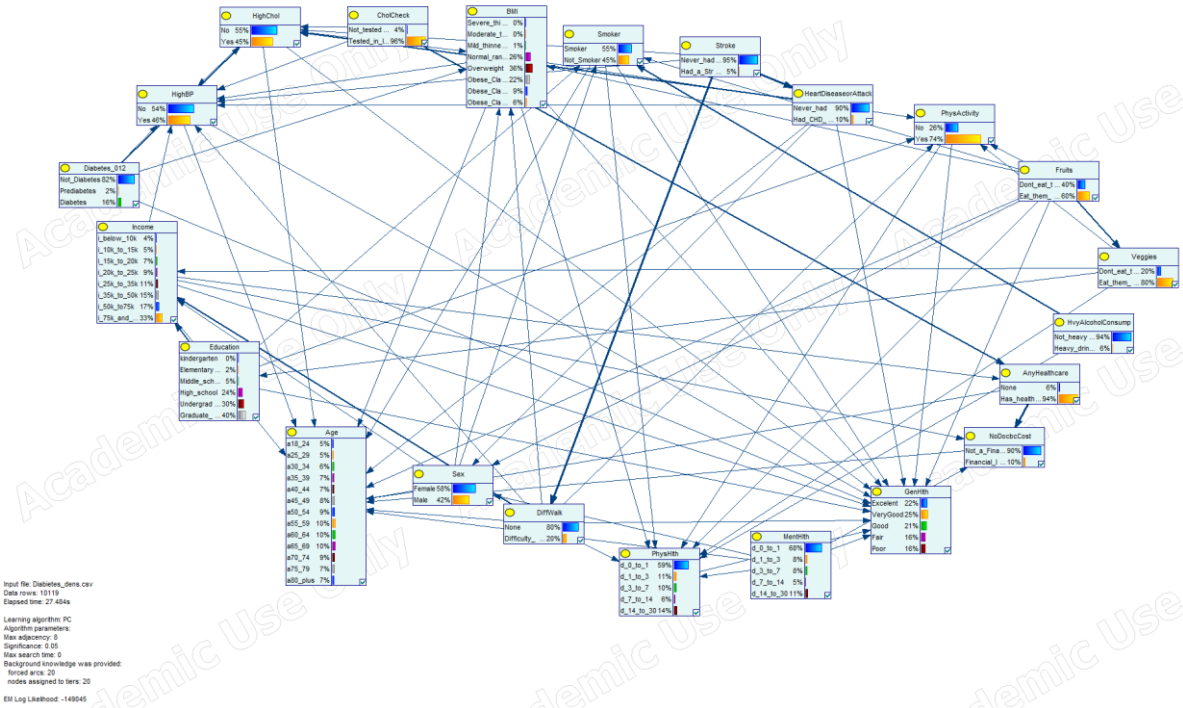


*Rysunek 8 Krzywa ROC dla modelu **Bayesian Search** dla odpowiedzi **Stan_przed_cukrzycowy**.*



*Rysunek 9 Krzywa ROC dla modelu **Bayesian Search** dla odpowiedzi **Cukrzyk**.*

- PC



*Rysunek 10 Struktura przyczynowo- skutkowa **PC** z węzłami w wersji prezentacji z rozkładami brzegowymi.*

Basic statistics:

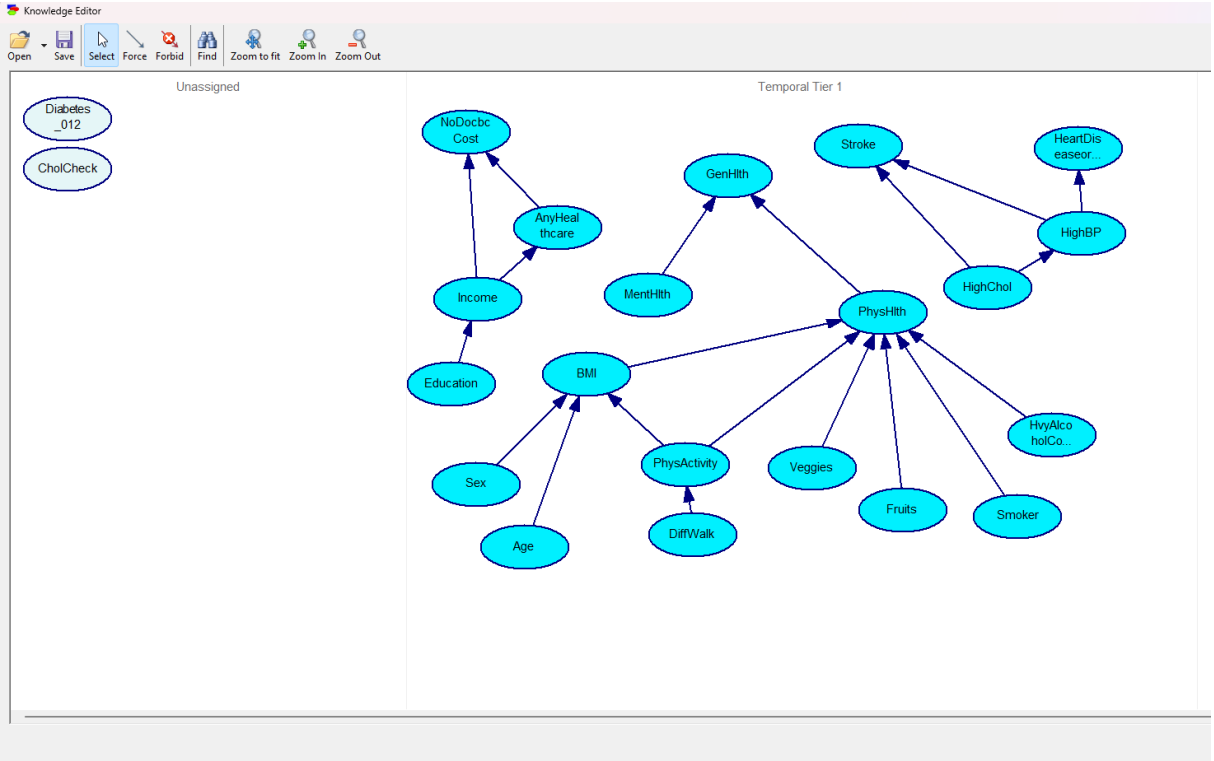
Node count: 22
Avg indegree: 3.409
Max indegree: 13
Avg outcomes: 3.682
Max outcomes: 13

Objects in the network:

Object	Count	States	Parameters / Independent
Nodes	22	81	20588960 / 16729334
Chance - General	22	81	20588960 / 16729334
Arcs	75		
Text boxes	1		

Rysunek 11 Podstawowe Informacje o modelu PC.

Liczba węzłów: 22
Średnia liczba rodziców węzła (średnia liczba stopni): 3,409
Maksymalna liczba rodziców węzła (maks. Liczba stopni): 13
Średnia liczba wyników węzłów: 3,6282
Maksymalna liczba wyników węzłów: 13
Ilość krawędzi: 75



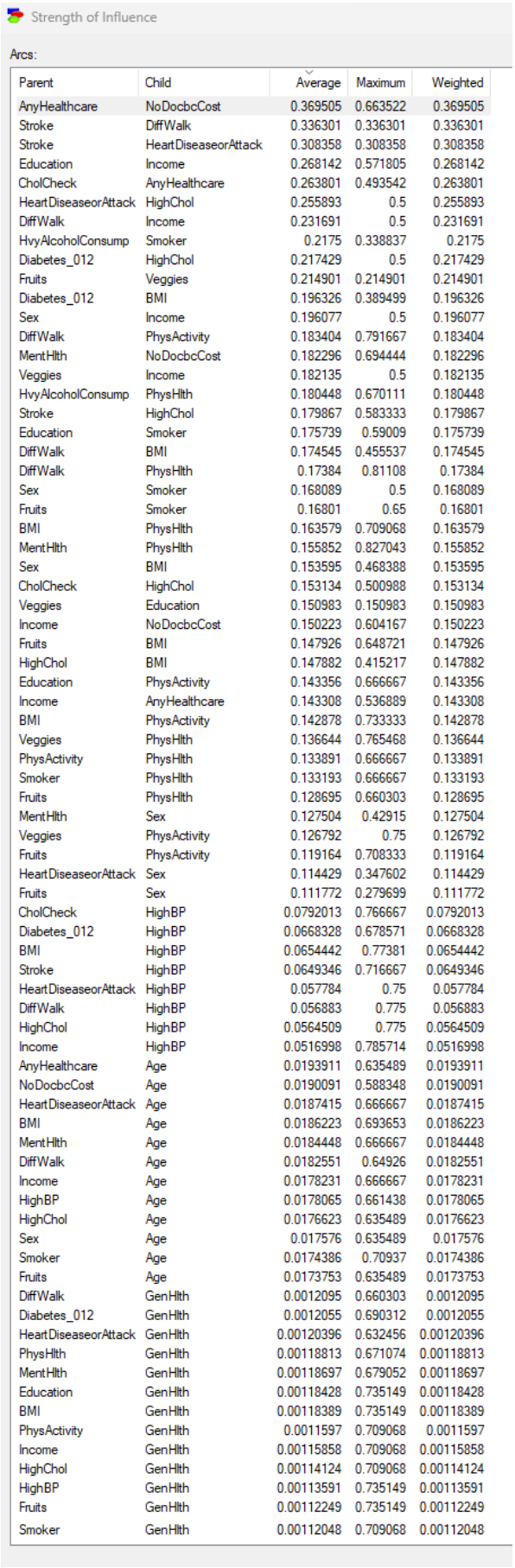
*Rysunek 12 Background knowledge dla modelu **PC**.*

Powyżej znajduje się schemat **Background knowledge** dla modelu **PC**. W *Temporal Tier 1* umieszczono większość czynników, które zostały podzielone na 3 grupy

-Grupa NoDocbcCost: *Edukacja* wpływa na *Zarobki*, które mają wpływ bezpośredni wpływ na *Ubezpieczenie zdrowotne* oraz na *Brak wizyty z względu na koszta*. Dodatkowo *Ubezpieczenie zdrowotne* również ma wpływ na *Brak wizyty z względu na koszta* (NoDocbcCost).

-Grupa GenHlth: *Problemy z poruszaniem* się mają wpływ na *Aktywność fizyczną*. *Aktywność Fizyczna* ma wpływ na *BMI* oraz *Zdrowie Fizyczne. Wiek* i *Płeć również mają wpływ na BMI. Warzywa, Owoce, Palenie, BMI* oraz *Alkoholizm* mają wpływ na *Zdrowie fizyczne. Zdrowie Fizyczne* oraz *Zdrowie Psychiczne* mają wpływ na *Generalne zdrowie* (GenHlth).

-Grupa Stroke: *Wysoki Cholesterol* ma wpływ na *Udar* oraz *Wysokie Ciśnienie*. *Wysokie Ciśnienie* ma wpływ na *Choroby Serca oraz na Udar* (Stroke).



Parent	Child	Average	Maximum	Weighted
AnyHealthcare	NoDocbcCost	0.369505	0.663522	0.369505
Stroke	DiffWalk	0.336301	0.336301	0.336301
Stroke	HeartDiseaseorAttack	0.308358	0.308358	0.308358
Education	Income	0.268142	0.571805	0.268142
CholCheck	AnyHealthcare	0.263801	0.493542	0.263801
HeartDiseaseorAttack	HighChol	0.255893	0.5	0.255893
DiffWalk	Income	0.231691	0.5	0.231691
HvyAlcoholConsump	Smoker	0.2175	0.338837	0.2175
Diabetes_012	HighChol	0.217429	0.5	0.217429
Fruits	Veggies	0.214901	0.214901	0.214901
Diabetes_012	BMI	0.196326	0.389499	0.196326
Sex	Income	0.196077	0.5	0.196077
DiffWalk	PhysActivity	0.183404	0.791667	0.183404
MentHlth	NoDocbcCost	0.182296	0.694444	0.182296
Veggies	Income	0.182135	0.5	0.182135
HvyAlcoholConsump	PhysHlth	0.180448	0.670111	0.180448
Stroke	HighChol	0.179867	0.583333	0.179867
Education	Smoker	0.175739	0.59009	0.175739
DiffWalk	BMI	0.174545	0.455537	0.174545
DiffWalk	PhysHlth	0.17384	0.81108	0.17384
Sex	Smoker	0.168089	0.5	0.168089
Fruits	Smoker	0.16801	0.65	0.16801
BMI	PhysHlth	0.163579	0.709068	0.163579
MentHlth	PhysHlth	0.155852	0.827043	0.155852
Sex	BMI	0.153595	0.468388	0.153595
CholCheck	HighChol	0.153134	0.500988	0.153134
Veggies	Education	0.150983	0.150983	0.150983
Income	NoDocbcCost	0.150223	0.604167	0.150223
Fruits	BMI	0.147926	0.648721	0.147926
HighChol	BMI	0.147882	0.415217	0.147882
Education	PhysActivity	0.143356	0.666667	0.143356
Income	AnyHealthcare	0.143308	0.536889	0.143308
BMI	PhysActivity	0.142878	0.733333	0.142878
Veggies	PhysHlth	0.136644	0.765468	0.136644
PhysActivity	PhysHlth	0.133891	0.666667	0.133891
Smoker	PhysHlth	0.133193	0.666667	0.133193
Fruits	PhysHlth	0.128695	0.660303	0.128695
MentHlth	Sex	0.127504	0.42915	0.127504
Veggies	PhysActivity	0.126792	0.75	0.126792
Fruits	PhysActivity	0.119164	0.708333	0.119164
HeartDiseaseorAttack	Sex	0.114429	0.347602	0.114429
Fruits	Sex	0.111772	0.279699	0.111772
CholCheck	HighBP	0.0792013	0.766667	0.0792013
Diabetes_012	HighBP	0.0668328	0.678571	0.0668328
BMI	HighBP	0.0654442	0.77381	0.0654442
Stroke	HighBP	0.0649346	0.716667	0.0649346
HeartDiseaseorAttack	HighBP	0.057784	0.75	0.057784
DiffWalk	HighBP	0.056883	0.775	0.056883
HighChol	HighBP	0.0564509	0.775	0.0564509
Income	HighBP	0.0516998	0.785714	0.0516998
AnyHealthcare	Age	0.0193911	0.635489	0.0193911
NoDocbcCost	Age	0.0190091	0.588348	0.0190091
HeartDiseaseorAttack	Age	0.0187415	0.666667	0.0187415
BMI	Age	0.0186223	0.693653	0.0186223
MentHlth	Age	0.0184448	0.666667	0.0184448
DiffWalk	Age	0.0182551	0.64926	0.0182551
Income	Age	0.0178231	0.666667	0.0178231
HighBP	Age	0.0178065	0.661438	0.0178065
HighChol	Age	0.0176623	0.635489	0.0176623
Sex	Age	0.017576	0.635489	0.017576
Smoker	Age	0.0174386	0.70937	0.0174386
Fruits	Age	0.0173753	0.635489	0.0173753
DiffWalk	GenHlth	0.0012095	0.660303	0.0012095
Diabetes_012	GenHlth	0.0012055	0.690312	0.0012055
HeartDiseaseorAttack	GenHlth	0.00120396	0.632456	0.00120396
PhysHlth	GenHlth	0.00118813	0.671074	0.00118813
MentHlth	GenHlth	0.00118697	0.679052	0.00118697
Education	GenHlth	0.00118428	0.735149	0.00118428
BMI	GenHlth	0.00118389	0.735149	0.00118389
PhysActivity	GenHlth	0.0011597	0.709068	0.0011597
Income	GenHlth	0.00115858	0.709068	0.00115858
HighChol	GenHlth	0.00114124	0.709068	0.00114124
HighBP	GenHlth	0.00113591	0.735149	0.00113591
Fruits	GenHlth	0.00112249	0.735149	0.00112249
Smoker	GenHlth	0.00112048	0.709068	0.00112048

*Rysunek 13 Lista krawędzi modelu **PC** z najsilniejszą mocą wpływu.*

Powyżej przedstawiona lista krawędzi dla modelu **PC** wskazuje, że krawędź z najsilniejszą mocą wpływu to ta **AnyHealthcare – NoDocbcCost** (Ubezpieczenie zdrowotne – Brak wizyty z względu na koszty), ubezpieczenie zdrowotne zmniejsza bariery w dostępie do opieki zdrowotnej.

Kolejne krawędzie z wyraźną mocą to:

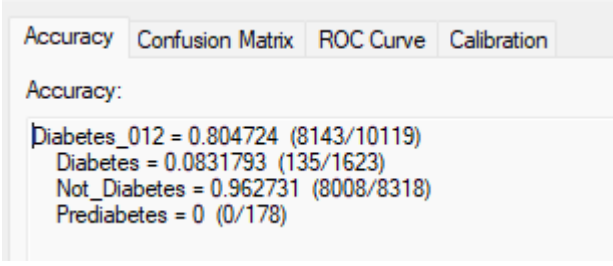
- **Stroke – DiffWalk** (Udar – Problemy z poruszaniem), gdzie osoby po udarze mają znaczne trudności w poruszaniu się.
- **Stroke – HeartDiseaseorAttack** (Udar – Choroby serca lub zawał), wskazując na silne powiązanie między przebytym udarem a ryzykiem chorób serca lub zawału.
- **Education – Income** (Edukacja – Dochód), gdzie wyższy poziom edukacji prowadzi do wyższych dochodów.
- **CholCheck – AnyHealthcare** (Badanie cholesterolu – Ubezpieczenie zdrowotne), co sugeruje, że osoby z ubezpieczeniem zdrowotnym częściej wykonują badania cholesterolu.
- **HeartDiseaseorAttack – HighChol** (Choroby serca lub zawał – Wysoki cholesterol), gdzie wysoki cholesterol jest silnie powiązany z chorobami serca i zawałami.

Dodatkowe istotne zależności to:

- **Diabetes_012 – HighChol** (Cukrzyca – Wysoki cholesterol), gdzie cukrzyca jest powiązana z wysokim cholesterolem.
- **Diabetes_012 – BMI** (Cukrzyca – Wskaźnik masy ciała), wskazując, że cukrzyca jest silnie powiązana z BMI.
- **DiffWalk – Income** (Problemy z poruszaniem – Dochód), gdzie trudności w poruszaniu się są związane z niższymi dochodami.
- **HvyAlcoholConsump – Smoker** (Nadmierne spożycie alkoholu – Palenie), wskazujące, że osoby pijące nadmierne ilości alkoholu często również palą.

Przy próbie uruchomienia diagnozy dla modelu PC program GeNIe Academic 4.1 przestał działać usuwając tym samym niezapisane postępy pracy. Przy ustawianiu properties dla *Age* oraz *GenHlt* program wyrzucał błąd „Out of memory”. Proces przeprowadzono kilkukrotnie bez sukcesu,

Walidacja za pomocą metody **Leave one out**. W przypadku modelu PC proces ten zajął bardzo długi czas ~25 minut!



*Rysunek 14 Lista Dokładności dla modelu **PC**.*

Jakość modelu **PC** wynosi ~0.804 czyli około 80,4%. Został on wyliczony na podstawie ilości poprawnie przewidzianych odpowiedzi czy osoba ma cukrzyce, czy jest w tanie przecukrzycowym lub nie ma cukrzycy. Model odgadł 8143 na 10119 przypadków. Jest to wynik najlepszy wśród badanych modeli. Przy sprawdzaniu jakości modelu zastosowano metodę **Leave one out.**

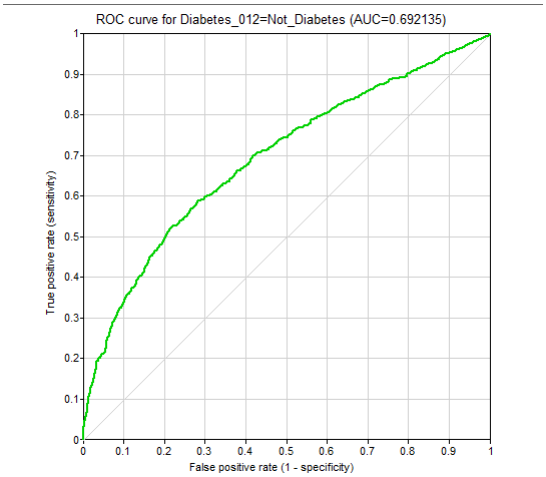
Accuracy | Confusion Matrix | ROC Curve | Calibration

Class node: Diabetes_012

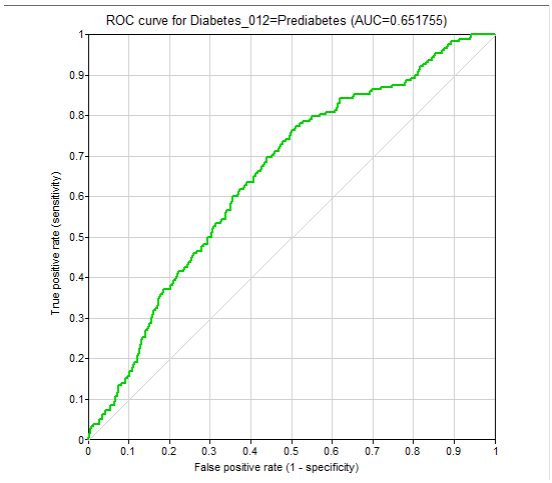
		Predicted		
		Diabetes	Not_Diabetes	Prediabetes
Actual	Diabetes	**135**	1487	1
	Not_Diabetes	307	**8008**	3
	Prediabetes	11	167	**0**

*Rysunek 15 Macierz Pomyłek dla modelu **PC**.*

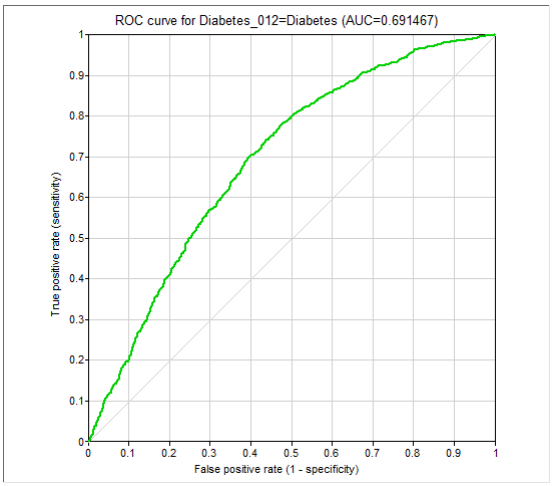
Z macierzy pomyłek dla modelu **PC** wynika, że dla wartości "*Not_Diabetes*" zostało poprawnie przewidzianych 8008 odpowiedzi, a błędnie 310. Dla odpowiedzi "*Prediabetes*" zostało poprawnie przewidzianych 0 odpowiedzi, a błędnie 178. Dla odpowiedzi "*Diabetes*" zostało poprawnie przewidzianych 135 odpowiedzi, a błędnie 1488. Łącznie model przewidział poprawnie 8143 wartości, a błędnie 1976 odpowiedzi.



*Rysunek 16 Krzywa ROC dla modelu **PC** dla odpowiedzi **Nie_Cukrzyk**.*

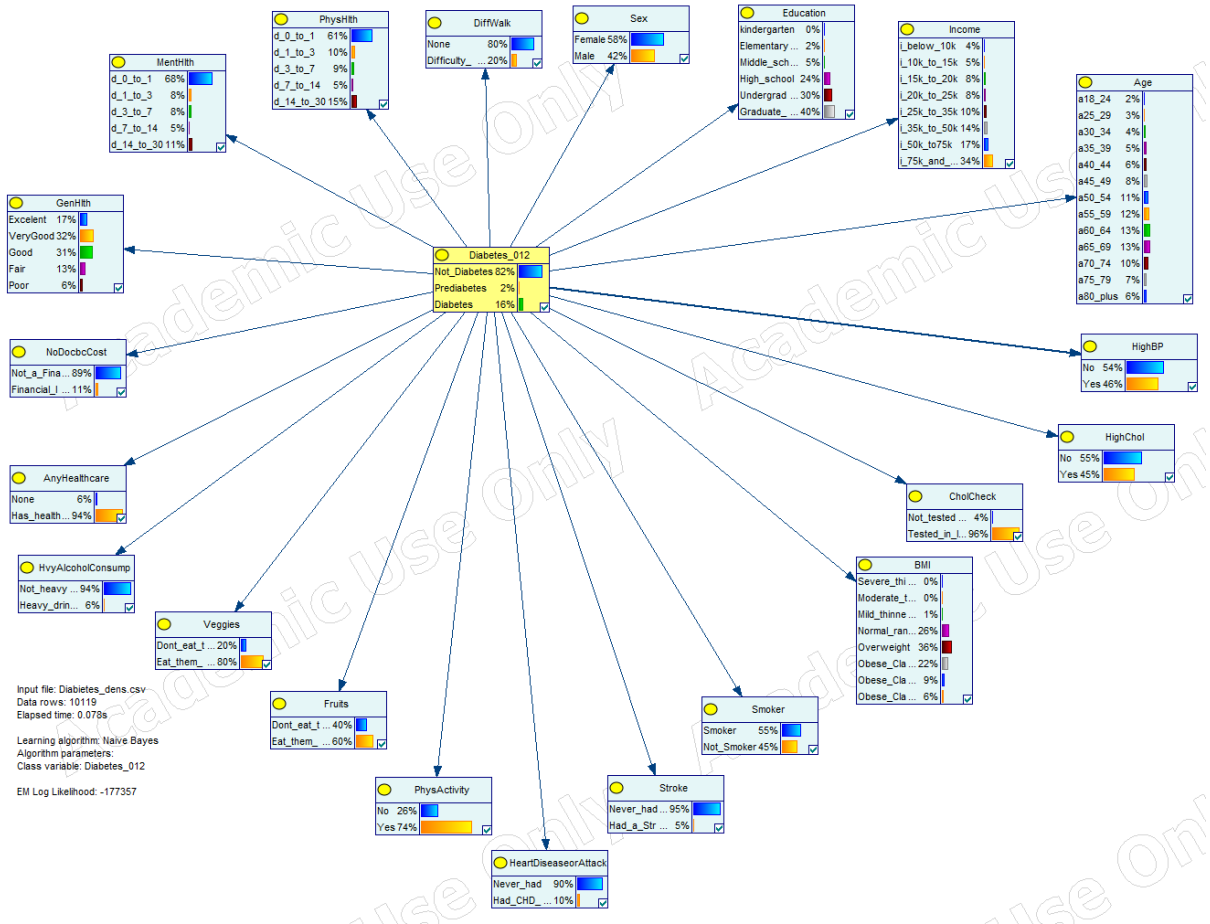


*Rysunek 17 Krzywa ROC dla modelu **PC** dla odpowiedzi **Stan_przed_Cukrzycowy**.*

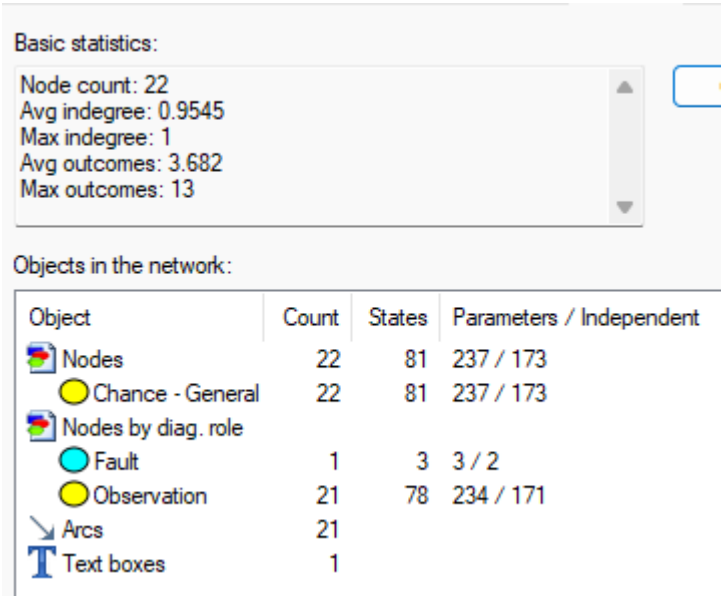


*Rysunek 18 Krzywa ROC dla modelu **PC** dla odpowiedzi **Cukrzyk**.*

- Naiwny Bayes



*Rysunek 19 Struktura przyczynowo- skutkowa **Naive Bayes** z węzłami w wersji prezentacji z rozkładami brzegowymi.*



*Rysunek 20 Pdostawowe informacje o modelu **Naive Bayes**.*

Liczba węzłów: 22
Średnia liczba rodziców węzła (średnia liczba stopni): 0,95
Maksymalna liczba rodziców węzła (maks. Liczba stopni): 1

Średnia liczba wyników węzłów: 3,682
Maksymalna liczba wyników węzłów: 13
Ilość krawędzi: 21

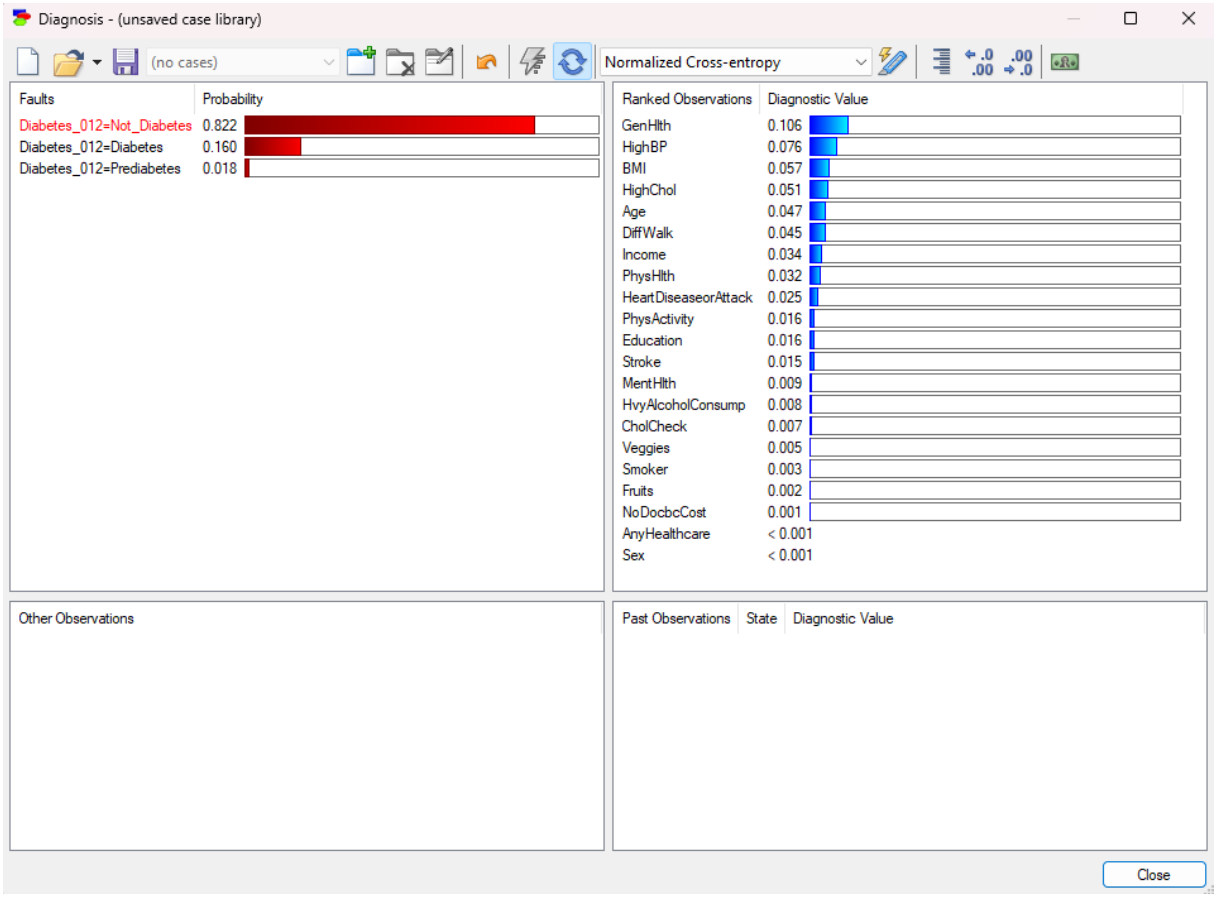
Strength of Influence

Arcs:

Parent	Child	Average	Maximum	Weighted
Diabetes_012	HighBP	0.238757	0.358135	0.238757
Diabetes_012	HighChol	0.190417	0.285626	0.190417
Diabetes_012	GenHlth	0.159554	0.235634	0.159554
Diabetes_012	DiffWalk	0.153416	0.230124	0.153416
Diabetes_012	BMI	0.128612	0.188614	0.128612
Diabetes_012	PhysHlth	0.123534	0.170203	0.123534
Diabetes_012	Income	0.109913	0.142715	0.109913
Diabetes_012	MentHlth	0.106074	0.152995	0.106074
Diabetes_012	PhysActivity	0.0977745	0.146662	0.0977745
Diabetes_012	HeartDiseaseorAttack	0.0918211	0.137732	0.0918211
Diabetes_012	Education	0.0878404	0.119475	0.0878404
Diabetes_012	Age	0.0814525	0.0955242	0.0814525
Diabetes_012	Fruits	0.0535079	0.0802618	0.0535079
Diabetes_012	Veggies	0.05022	0.0753301	0.05022
Diabetes_012	Stroke	0.0494421	0.0741631	0.0494421
Diabetes_012	Smoker	0.043976	0.065964	0.043976
Diabetes_012	Sex	0.0399316	0.0598974	0.0399316
Diabetes_012	HvyAlcoholConsump	0.0312273	0.0468409	0.0312273
Diabetes_012	CholCheck	0.0243848	0.0365772	0.0243848
Diabetes_012	NoDocbcCost	0.0216821	0.0325232	0.0216821
Diabetes_012	AnyHealthcare	0.0165014	0.0247521	0.0165014

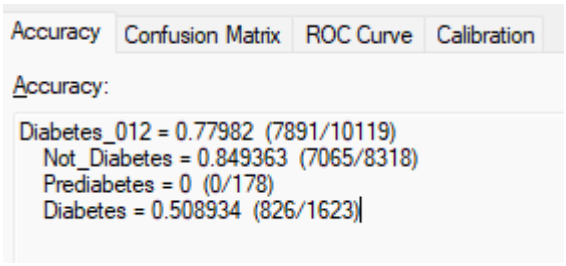
*Rysunek 21 Lista krawędzi modelu **Naive Bayes** z najsilniejszą mocą wpływu.*

W modelu naiwnego Bayesa krawędź z najsilniejszą mocą wpływu to krawędź Diabetes_012 - HighBP (problemy z wysokim ciśnieniem krwi). Kolejnymi dwiema kluczowymi krawędziami są Diabetes_012 - HighChol (problemy z wysokim choresterolem), Diabetes_012 – GenHlth (ocena samodzielna stanu zdrowia) oraz Diabetes_012 – DiffWalk (problemy z poruszaniem się). Głównie od tych czterech połączeń zależy przewidywanie czy osoba może mieć cukrzyce oraz czy jest na to duże prawdopodobieństwo. Reszta połączeń ma tutaj mniejsze znaczenie od wcześniej wymienionych. Najmniej znaczącym połączeniem jest krawędź HeartDisease -AnyHealthcare (jakiekolwiek ubezpieczenie zdrowotne).



*Rysunek 22 Lista zmiennych modelu **Naive Bayes**, które mają najwyższą wartość diagnostyczną.*

Powyżej znajduje się lista zmiennych modelu **naiwnego Bayesa**, które mają najwyższą wartość diagnostyczną w celu różnicowania zmiennej, która reprezentuje klasę *Diabetes_012* (cukrzyca). Najwyższą wartość diagnostyczną uzyskała zmienna *GenHlth* (ocena samodzielna stanu zdrowia) uzyskując wynik 0.106. Kolejnymi zmiennymi o wysokiej wartości diagnostycznej są: *HighBP* (problemy z wysokim ciśnieniem krwi) wynik: 0.076, *BMI* wynik 0.057, *HighChol* (wysoki cholesterol) z wynikiem 0.051, *Age* (kategoria wiekowa) wynik 0.047, *DiffWalk* (trudność w poruszaniu się) wynik 0.045. Reszta zmiennych ma mniejszą wartość diagnostyczną od wymienionych wyżej zmiennych. Najmniejszą wartość ma *Sex* (płeć) oraz *AnyHealthcare* (jakiekolwiek ubezpieczenie zdrowotne) z wynikiem mniejszym niż 0.001.



*Rysunek 23 Lista Dokładności dla modelu **Naive Bayes**.*

Jakość modelu **naiwnego Bayesa** wynosi ~0.779 czyli około 77,9%. Został on wyliczony na podstawie ilości poprawnie przewidzianych odpowiedzi czy osoba ma cukrzyce,

czy jest w tanie przecukrzycowym lub nie ma cukrzycy. Model odgadł 7891 na 10119 przypadków. Przy sprawdzaniu jakości modelu zastosowano metodę **Leave one out.**

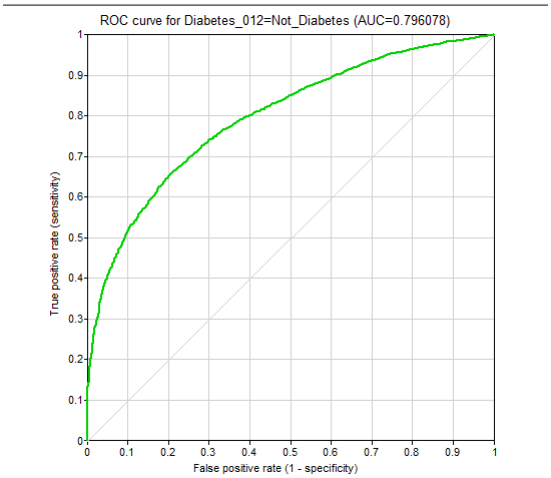
Accuracy | Confusion Matrix | ROC Curve | Calibration

Class node: Diabetes_012

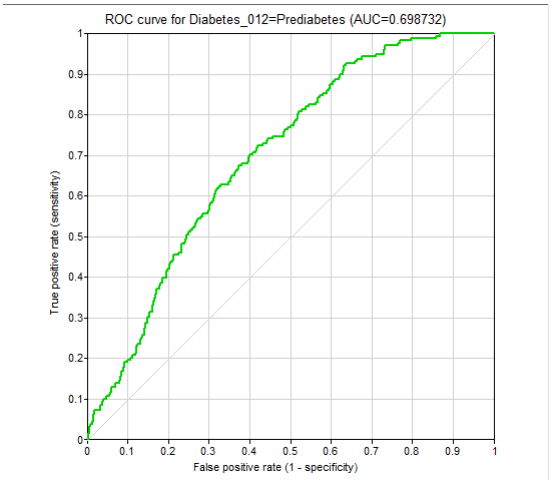
		Predicted		
		Not_Diabetes	Prediabetes	Diabetes
Actual	Not_Diabetes	**7065**	5	1248
	Prediabetes	103	**0**	75
	Diabetes	796	1	**826**

*Rysunek 24 Macierz pomyłek modelu **Naive Bayes**.*

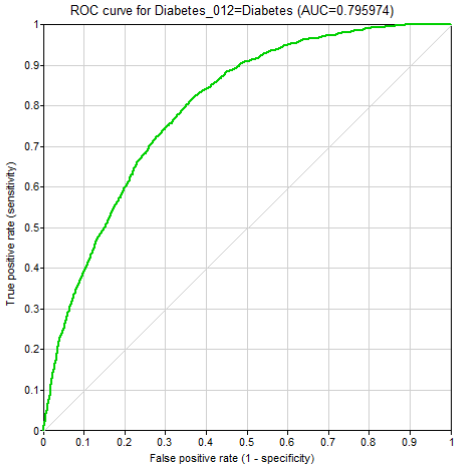
Z macierzy pomyłek dla modelu **Naive Bayes** wynika, że dla wartości "*Not_Diabetes*" zostało poprawnie przewidzianych 7065 odpowiedzi, a błędnie 1253. Dla odpowiedzi "*Prediabetes*" zostało poprawnie przewidzianych 0 odpowiedzi, a błędnie 178. Dla odpowiedzi "*Diabetes*" zostało poprawnie przewidzianych 826 odpowiedzi, a błędnie 797. Łącznie model przewidział poprawnie 7891 wartości, a błędnie 2228 odpowiedzi.



*Rysunek 25 Krzywa ROC dla modelu **Naive Bayes** dla odpowiedzi **Nie_Cukrzyk**.*

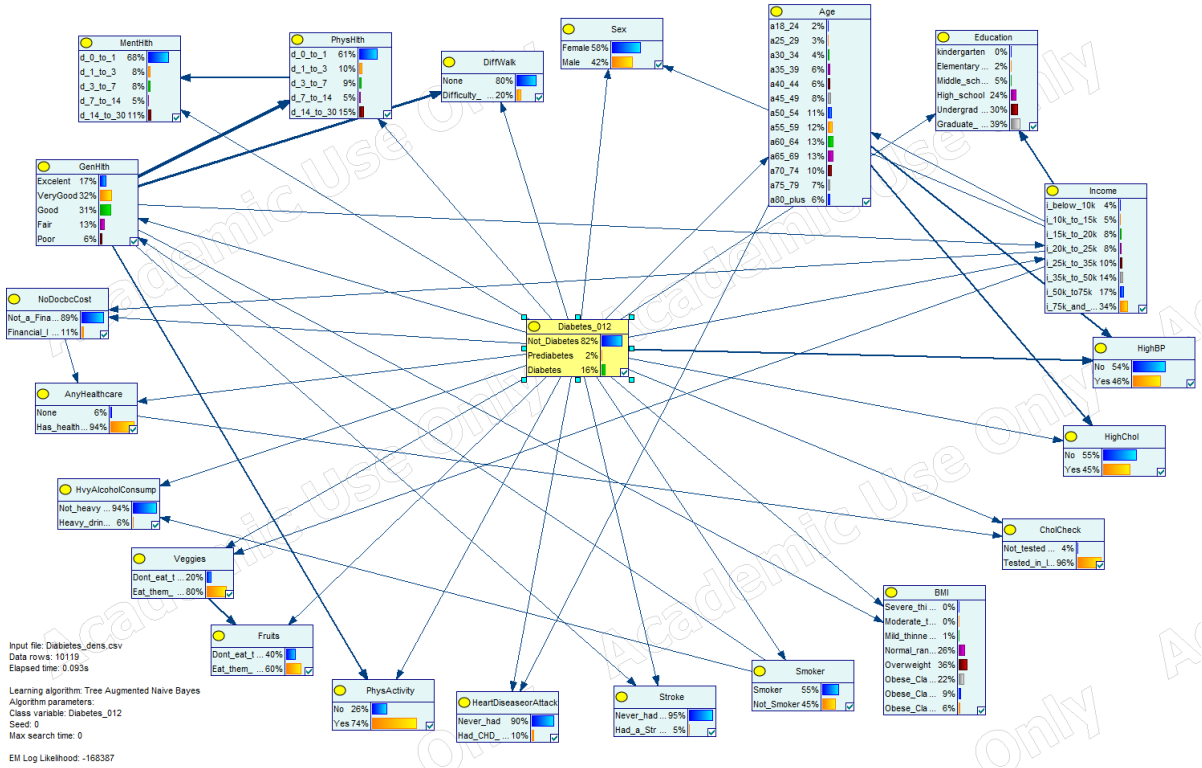


*Rysunek 26 Krzywa ROC dla modelu **Naive Bayes** dla odpowiedzi **Stan_przed_Cukrzycowy**.*



*Rysunek 27 Krzywa ROC dla modelu **Naive Bayes** dla odpowiedzi **Cukrzyk**.*

- TAN- Tree Augmented Naive Bayes



*Rysunek 28 Struktura przyczynowo- skutkowa **Tree Augmented Naive Bayes** z węzłami w wersji prezentacji z rozkładami brzegowymi.*

Basic statistics:

Node count: 22
Avg indegree: 1.864
Max indegree: 2
Avg outcomes: 3.682
Max outcomes: 13

Objects in the network:

Object	Count	States	Parameters / Independent
Nodes	22	81	1401 / 1025
Chance - General	22	81	1401 / 1025
Arcs	41		
Text boxes	1		

*Rysunek 29 Podstawowe dane modelu **Tree Augmented Naive Bayes**.*

Liczba węzłów: 22
Średnia liczba rodziców węzła (średnia liczba stopni): 1,864
Maksymalna liczba rodziców węzła (maks. Liczba stopni): 2
Średnia liczba wyników węzłów: 3,682
Maksymalna liczba wyników węzłów: 13
Ilość krawędzi: 41

Strength of Influence

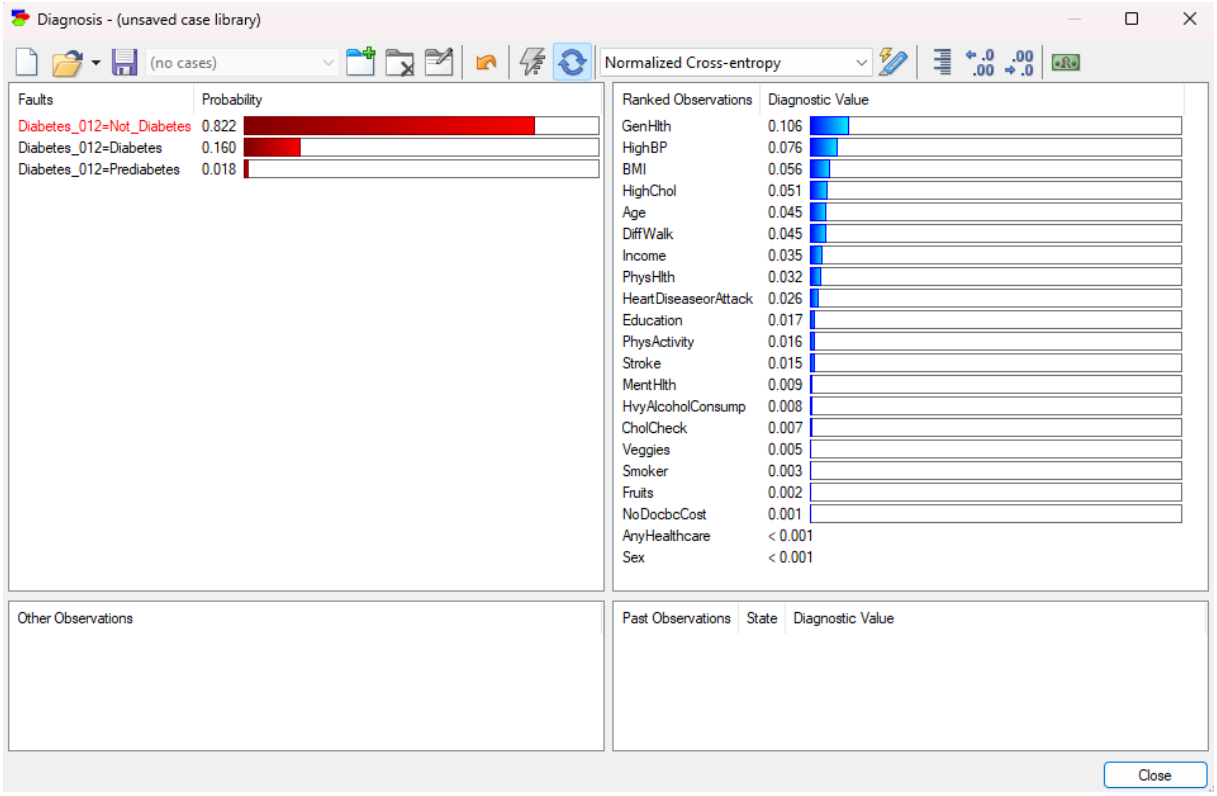
Arcs:

Parent	Child	Average	Maximum	Weighted
GenHlth	PhysHlth	0.411465	0.801806	0.411465
GenHlth	DiffWalk	0.378357	0.865946	0.378357
Veggies	Fruits	0.270677	0.346484	0.270677
PhysHlth	MentHlth	0.240161	0.447934	0.240161
Income	Education	0.219532	0.474169	0.219532
GenHlth	PhysActivity	0.219359	0.47415	0.219359
Age	HighBP	0.21866	0.654525	0.21866
Age	HighChol	0.216408	0.657609	0.216408
Diabetes_012	HighBP	0.203157	0.482073	0.203157
GenHlth	Income	0.193189	0.358511	0.193189
Diabetes_012	HighChol	0.191752	0.472436	0.191752
NoDocbcCost	AnyHealthcare	0.185185	0.212675	0.185185
AnyHealthcare	CholCheck	0.164813	0.282807	0.164813
Diabetes_012	GenHlth	0.164794	0.263236	0.164794
Diabetes_012	BMI	0.158311	0.354412	0.158311
Income	Sex	0.151776	0.525901	0.151776
GenHlth	BMI	0.140497	0.355826	0.140497
Diabetes_012	MentHlth	0.137406	0.314966	0.137406
Diabetes_012	Age	0.130778	0.192665	0.130778
Diabetes_012	Income	0.130195	0.397691	0.130195
Income	NoDocbcCost	0.126928	0.330303	0.126928
Age	HeartDiseaseorAttack	0.125689	0.470588	0.125689
Income	Veggies	0.125037	0.455774	0.125037
Diabetes_012	HeartDiseaseorAttack	0.120467	0.497487	0.120467
Income	Age	0.11583	0.269467	0.11583
Diabetes_012	Education	0.104924	0.241347	0.104924
GenHlth	Stroke	0.0982883	0.305294	0.0982883
Diabetes_012	PhysActivity	0.0966738	0.45877	0.0966738
Diabetes_012	Sex	0.0870563	0.277828	0.0870563
Diabetes_012	CholCheck	0.0849468	0.222423	0.0849468
Smoker	GenHlth	0.0738208	0.0840875	0.0738208
Diabetes_012	DiffWalk	0.0714645	0.180388	0.0714645
Diabetes_012	Veggies	0.0633397	0.289279	0.0633397
Diabetes_012	PhysHlth	0.0609666	0.174602	0.0609666
Diabetes_012	Stroke	0.0565798	0.178726	0.0565798
Diabetes_012	NoDocbcCost	0.0512058	0.157606	0.0512058
Smoker	HvyAlcoholConsump	0.0466322	0.0618602	0.0466322
Diabetes_012	Fruits	0.0446294	0.0865008	0.0446294
Diabetes_012	Smoker	0.043976	0.065964	0.043976
Diabetes_012	AnyHealthcare	0.0370555	0.0883275	0.0370555
Diabetes_012	HvyAlcoholConsump	0.0335216	0.0662734	0.0335216

*Rysunek 30 Lista krawędzi modelu **Tree Augmented Naive Bayes** z najsilniejszą mocą wpływu.*

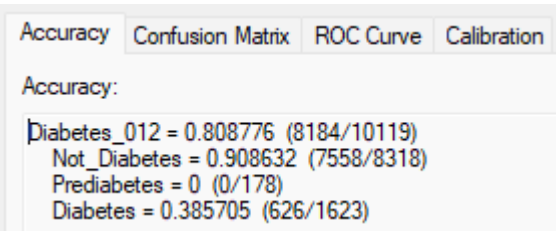
Model **Tree Augmented Naive Bayes** zawiera znacznie więcej krawędzi niż model naiwnego Bayesa. Krawędzią z najsilniejszą mocą wpływu to jest krawędź **GenHlth - PhysHlth** (generalna opinia o zdrowiu i zdrowie fizyczne ostatnich 30 dniach). Kolejne krawędzie z bardzo dużą mocą wpływu to **GenHlth - DiffWalk** (generalna opinia o zdrowiu i problemy z poruszaniem się), **Veggies - Fruits** (spożywanie codzienne warzyw oraz spożywanie codzienne owoców), **PhysHlth - MentHlth** (zdrowie fizyczne w ostatnich 30

dniach i zdrowie psychiczne w ostatnich 30 dniach). Głównie od tych połączeń zależy przewidywanie czy osoba może mieć cukrzyce, czy jest na to duże prawdopodobieństwo. Reszta połączeń ma tutaj mniejsze znaczenie od wcześniej wymienionych. Najmniej znaczącym połączeniem jest krawędź **Diaetes_012 - HvyAlcoholConsump** (cukrzyca i spożycie alkoholu).



*Rysunek 31 Lista zmiennych modelu **Tree Augmented Naive Bayes**, które mają najwyższą wartość diagnostyczną.*

Powyżej znajduje się lista zmiennych modelu Tree Augmented Naive Bayes, które mają najwyższą wartość diagnostyczną w celu różnicowania zmiennej, która reprezentuje klasę Diabetes_012 (cukrzyca). Jest ona bardzo podobna do listy zmiennych z modelu naiwnego Bayesa i modelu Augmented Naive Bayes. Najwyższą wartość diagnostyczną uzyskała zmienna GenHlth (ocena samodzielna stanu zdrowia) uzyskując wynik 0.106. Kolejnymi zmiennymi o wysokiej wartości diagnostycznej są: HighBP (wysokie ciśnienie krwi) wynik: 0.076, BMI wynik 0.056, HighChol (wysoki cholesterol) z wynikiem 0.051, Age (kategoria wiekowa) wynik 0.045, DifWalk (problemy z poruszaniem się) wynik 0.045. Reszta zmiennych ma mniejszą wartość diagnostyczną od wymienionych wyżej zmiennych. Najmniejszą wartość ma Sex (płeć) oraz AnyHealthcare (jakiekolwiek ubezpieczenie zdrowotne) z wynikiem mniejszym niż 0.001.



*Rysunek 32 Lista Dokładności dla modelu **Tree Augmented Naive Bayes**.*

Jakość modelu Tree Augmented Naive Bayes wynosi ~0.808 czyli około 80,8%. Został on wyliczony na podstawie ilości poprawnie przewidzianych odpowiedzi czy osoba ma cukrzyce, czy jest w tanie przecukrzycowym lub nie ma cukrzycy. Model odgadł 8148 na 10119 przypadków. Jest to wynik trochę słabszy niż uzyskał model ANB. Przy sprawdzaniu jakości modelu zastosowano metodę **Leave one out**.

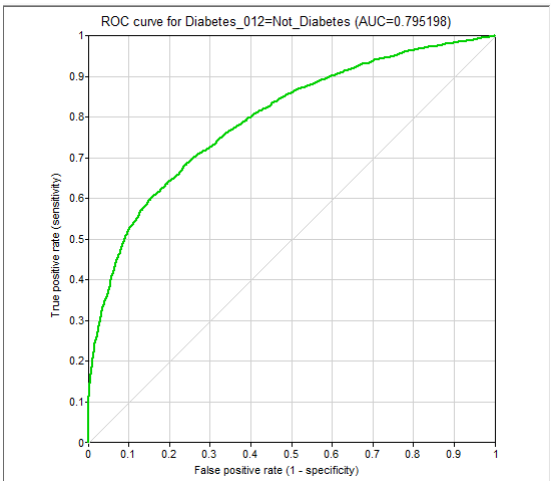
Accuracy | Confusion Matrix | ROC Curve | Calibration

Class node: Diabetes_012

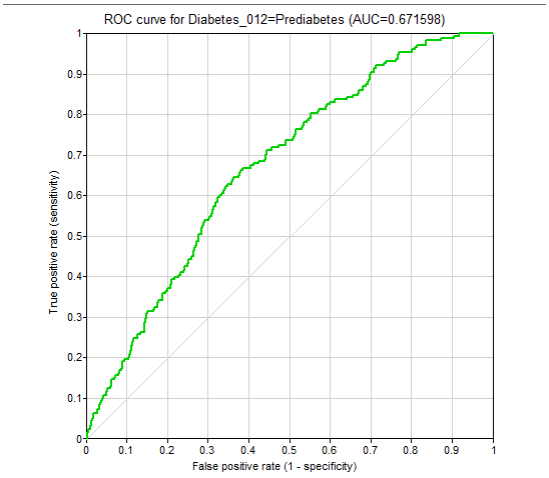
		Predicted		
		Not_Diabetes	Prediabetes	Diabetes
Actual	Not_Diabetes	**7558**	12	748
	Prediabetes	135	**0**	43
	Diabetes	987	10	**626**

*Rysunek 33 Macierz pomyłek modelu **Tree Augmented Naive Bayes**.*

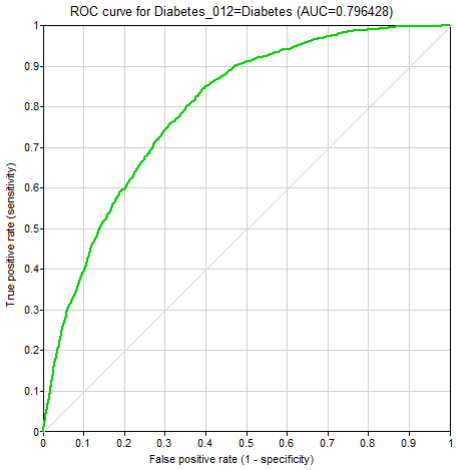
Z macierzy pomyłek dla modelu **Tree Augmented Naive Bayes** wynika, że dla wartości "*Not_Diabetes*" zostało poprawnie przewidzianych 7558 odpowiedzi, a błędnie 760. Dla odpowiedzi "*Prediabetes*" zostało poprawnie przewidzianych 0 odpowiedzi, a błędnie 178. Dla odpowiedzi "*Diabetes*" zostało poprawnie przewidzianych 626 odpowiedzi, a błędnie 997. Łącznie model przewidział poprawnie 8184 wartości, a błędnie 1935 odpowiedzi.



*Rysunek 34 Krzywa ROC dla modelu **Tree Augmented Naive Bayes** dla odpowiedzi **Nie_Cukrzyk**.*

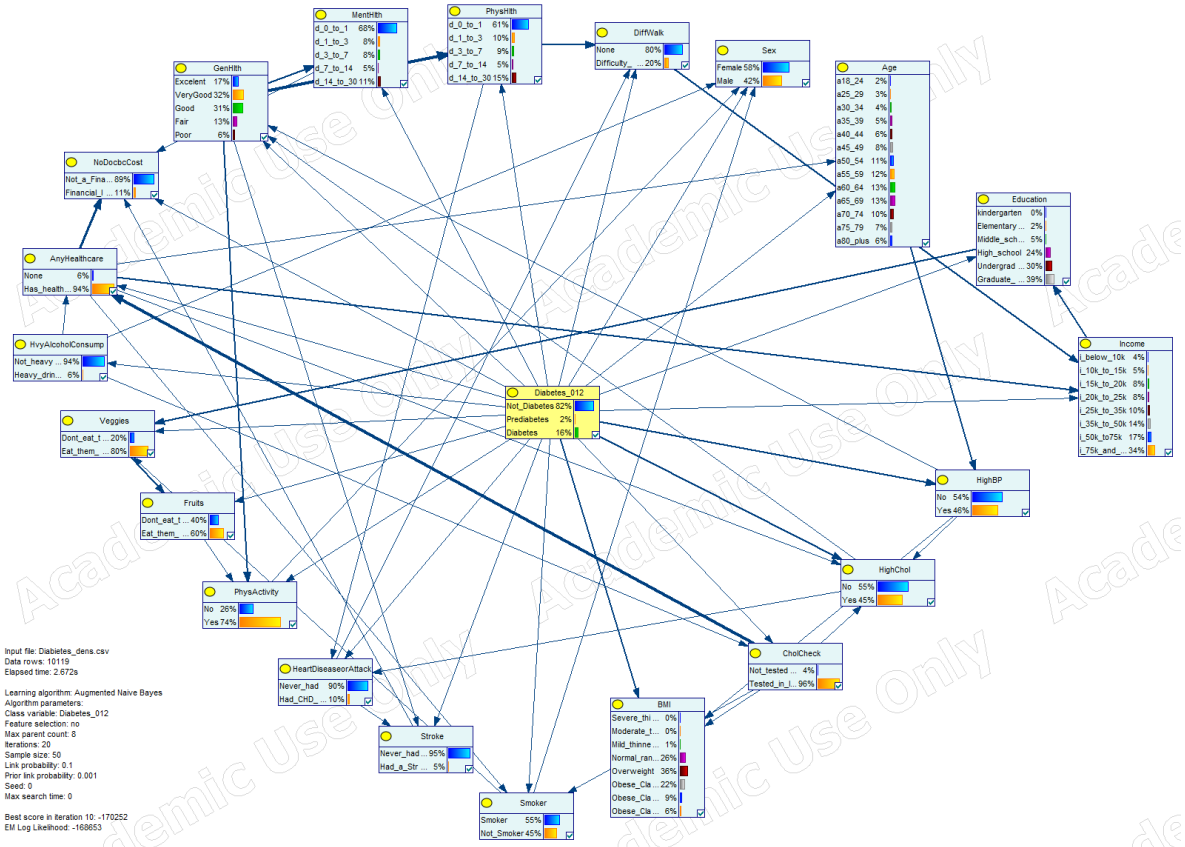


*Rysunek 35 Krzywa ROC dla modelu **Tree Augmented Naive Bayes** dla odpowiedzi **Stan_przed_Cukrzycowy**.*

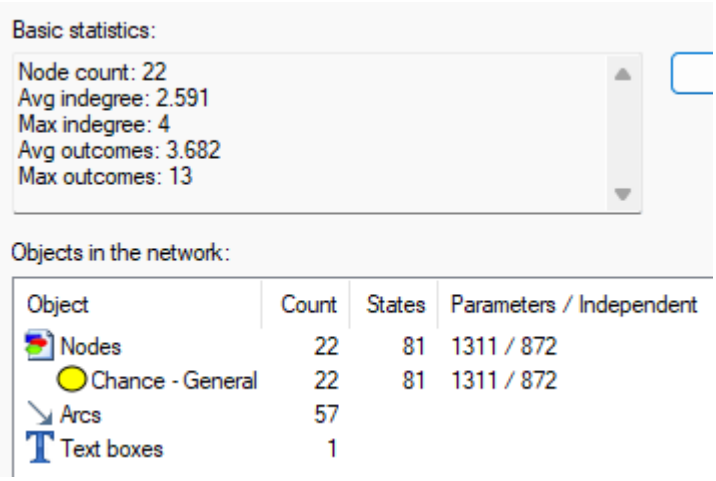


*Rysunek 36 Krzywa ROC dla modelu **Tree Augmented Naive Bayes** dla odpowiedzi **Cukrzyk**.*

- ANB- Augmented Naive Bayes:

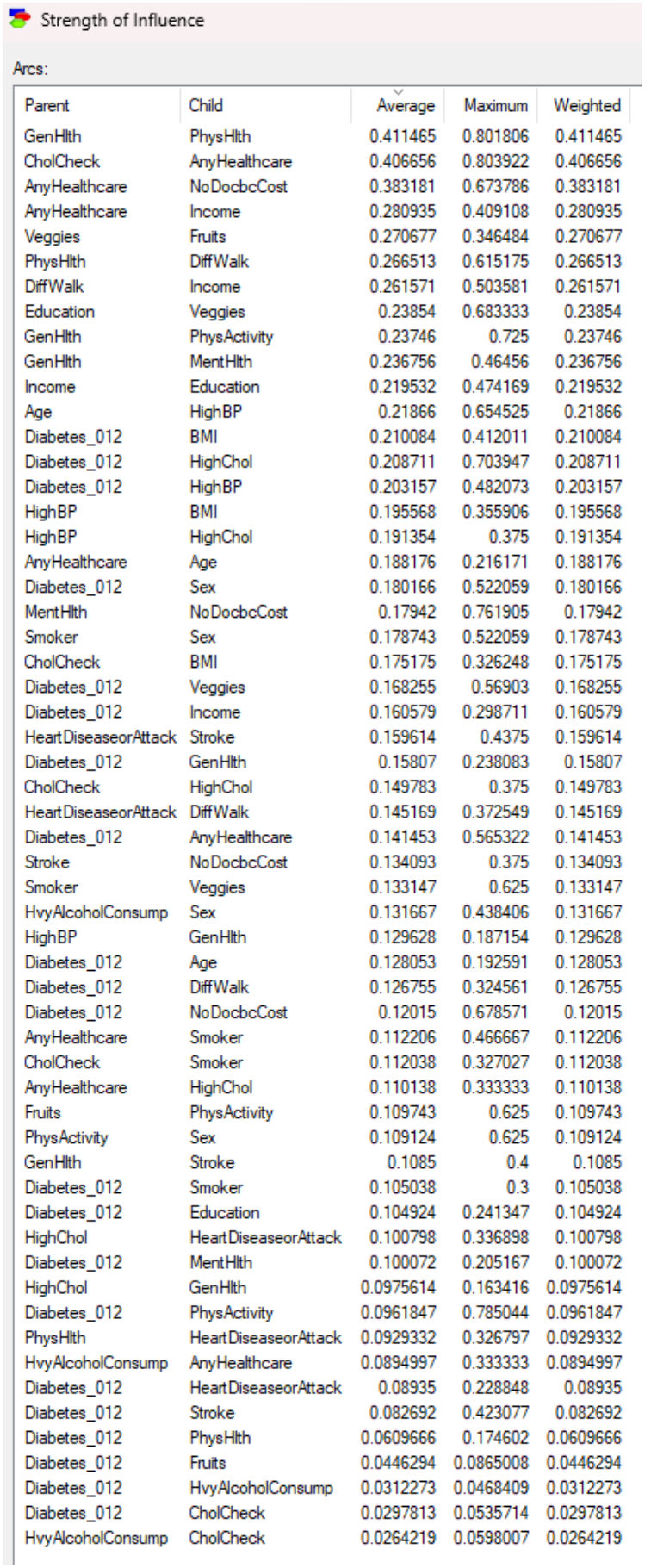


*Rysunek 37 Struktura przyczynowo- skutkowa **Augmented Naive Bayes** z węzłami w wersji prezentacji z rozkładami brzegowymi.*



*Rysunek 38 Podstawowe informacje o modelu **Augmented Naive Bayes**.*

Liczba węzłów: 22
Średnia liczba rodziców węzła (średnia liczba stopni): 2,591
Maksymalna liczba rodziców węzła (maks. Liczba stopni): 4
Średnia liczba wyników węzłów: 3,682
Maksymalna liczba wyników węzłów: 13
Ilość krawędzi: 57



Parent	Child	Average	Maximum	Weighted
GenHlth	PhysHlth	0.411465	0.801806	0.411465
CholCheck	AnyHealthcare	0.406656	0.803922	0.406656
AnyHealthcare	NoDocbcCost	0.383181	0.673786	0.383181
AnyHealthcare	Income	0.280935	0.409108	0.280935
Veggies	Fruits	0.270677	0.346484	0.270677
PhysHlth	DiffWalk	0.266513	0.615175	0.266513
DiffWalk	Income	0.261571	0.503581	0.261571
Education	Veggies	0.23854	0.683333	0.23854
GenHlth	PhysActivity	0.23746	0.725	0.23746
GenHlth	MentHlth	0.236756	0.46456	0.236756
Income	Education	0.219532	0.474169	0.219532
Age	HighBP	0.21866	0.654525	0.21866
Diabetes_012	BMI	0.210084	0.412011	0.210084
Diabetes_012	HighChol	0.208711	0.703947	0.208711
Diabetes_012	HighBP	0.203157	0.482073	0.203157
HighBP	BMI	0.195568	0.355906	0.195568
HighBP	HighChol	0.191354	0.375	0.191354
AnyHealthcare	Age	0.188176	0.216171	0.188176
Diabetes_012	Sex	0.180166	0.522059	0.180166
MentHlth	NoDocbcCost	0.17942	0.761905	0.17942
Smoker	Sex	0.178743	0.522059	0.178743
CholCheck	BMI	0.175175	0.326248	0.175175
Diabetes_012	Veggies	0.168255	0.56903	0.168255
Diabetes_012	Income	0.160579	0.298711	0.160579
HeartDiseaseorAttack	Stroke	0.159614	0.4375	0.159614
Diabetes_012	GenHlth	0.15807	0.238083	0.15807
CholCheck	HighChol	0.149783	0.375	0.149783
HeartDiseaseorAttack	DiffWalk	0.145169	0.372549	0.145169
Diabetes_012	AnyHealthcare	0.141453	0.565322	0.141453
Stroke	NoDocbcCost	0.134093	0.375	0.134093
Smoker	Veggies	0.133147	0.625	0.133147
HvyAlcoholConsump	Sex	0.131667	0.438406	0.131667
HighBP	GenHlth	0.129628	0.187154	0.129628
Diabetes_012	Age	0.128053	0.192591	0.128053
Diabetes_012	DiffWalk	0.126755	0.324561	0.126755
Diabetes_012	NoDocbcCost	0.12015	0.678571	0.12015
AnyHealthcare	Smoker	0.112206	0.466667	0.112206
CholCheck	Smoker	0.112038	0.327027	0.112038
AnyHealthcare	HighChol	0.110138	0.333333	0.110138
Fruits	PhysActivity	0.109743	0.625	0.109743
PhysActivity	Sex	0.109124	0.625	0.109124
GenHlth	Stroke	0.1085	0.4	0.1085
Diabetes_012	Smoker	0.105038	0.3	0.105038
Diabetes_012	Education	0.104924	0.241347	0.104924
HighChol	HeartDiseaseorAttack	0.100798	0.336898	0.100798
Diabetes_012	MentHlth	0.100072	0.205167	0.100072
HighChol	GenHlth	0.0975614	0.163416	0.0975614
Diabetes_012	PhysActivity	0.0961847	0.785044	0.0961847
PhysHlth	HeartDiseaseorAttack	0.0929332	0.326797	0.0929332
HvyAlcoholConsump	AnyHealthcare	0.0894997	0.333333	0.0894997
Diabetes_012	HeartDiseaseorAttack	0.08935	0.228848	0.08935
Diabetes_012	Stroke	0.082692	0.423077	0.082692
Diabetes_012	PhysHlth	0.0609666	0.174602	0.0609666
Diabetes_012	Fruits	0.0446294	0.0865008	0.0446294
Diabetes_012	HvyAlcoholConsump	0.0312273	0.0468409	0.0312273
Diabetes_012	CholCheck	0.0297813	0.0535714	0.0297813
HvyAlcoholConsump	CholCheck	0.0264219	0.0598007	0.0264219

*Rysunek 39 Lista krawędzi modelu **Augmented Naive Bayes** z najsilniejszą mocą wpływu.*

Powyżej przedstawiona lista krawędzi dla modelu **Augmented Naive Bayes** wskazuje, że krawędź z najsilniejszą mocą wpływu to ta **GenHlth – PhysHlth** (Ogólny stan zdrowia – Zdrowie fizyczne). Ogólny stan zdrowia silnie wpływa na zdrowie fizyczne.

Kolejne krawędzie z wyraźną mocą to:

- **CholCheck – AnyHealthcare** (Badanie cholesterolu – Ubezpieczenie zdrowotne), gdzie posiadanie ubezpieczenia zdrowotnego zwiększa prawdopodobieństwo wykonywania badań cholesterolu.

- **AnyHealthcare – NoDocbcCost** (Ubezpieczenie zdrowotne – Brak wizyty z względu na koszty), gdzie ubezpieczenie zdrowotne zmniejsza bariery w dostępie do opieki zdrowotnej.

- **AnyHealthcare – Income** (Ubezpieczenie zdrowotne – Dochód), gdzie wyższe dochody są powiązane z posiadaniem ubezpieczenia zdrowotnego.

- **Veggies – Fruits** (Warzywa – Owoce), co sugeruje, że osoby spożywające warzywa, również często spożywają owoce.

- **PhysHlth – DiffWalk** (Zdrowie fizyczne – Problemy z poruszaniem), gdzie gorsze zdrowie fizyczne zwiększa trudności w poruszaniu się.

Dodatkowe istotne zależności to:

- **DiffWalk – Income** (Problemy z poruszaniem – Dochód), gdzie trudności w poruszaniu się są związane z niższymi dochodami.

- **Education – Veggies** (Edukacja – Warzywa), gdzie wyższy poziom edukacji prowadzi do częstszego spożywania warzyw.

- **GenHlth – PhysActivity** (Ogólny stan zdrowia – Aktywność fizyczna), gdzie lepszy stan zdrowia jest powiązany z większą aktywnością fizyczną.

- **GenHlth – MentHlth** (Ogólny stan zdrowia – Zdrowie psychiczne), gdzie lepszy stan zdrowia ogólnego wpływa na lepsze zdrowie psychiczne.
- Inne zależności dotyczą bezpośrednio cukrzycy:

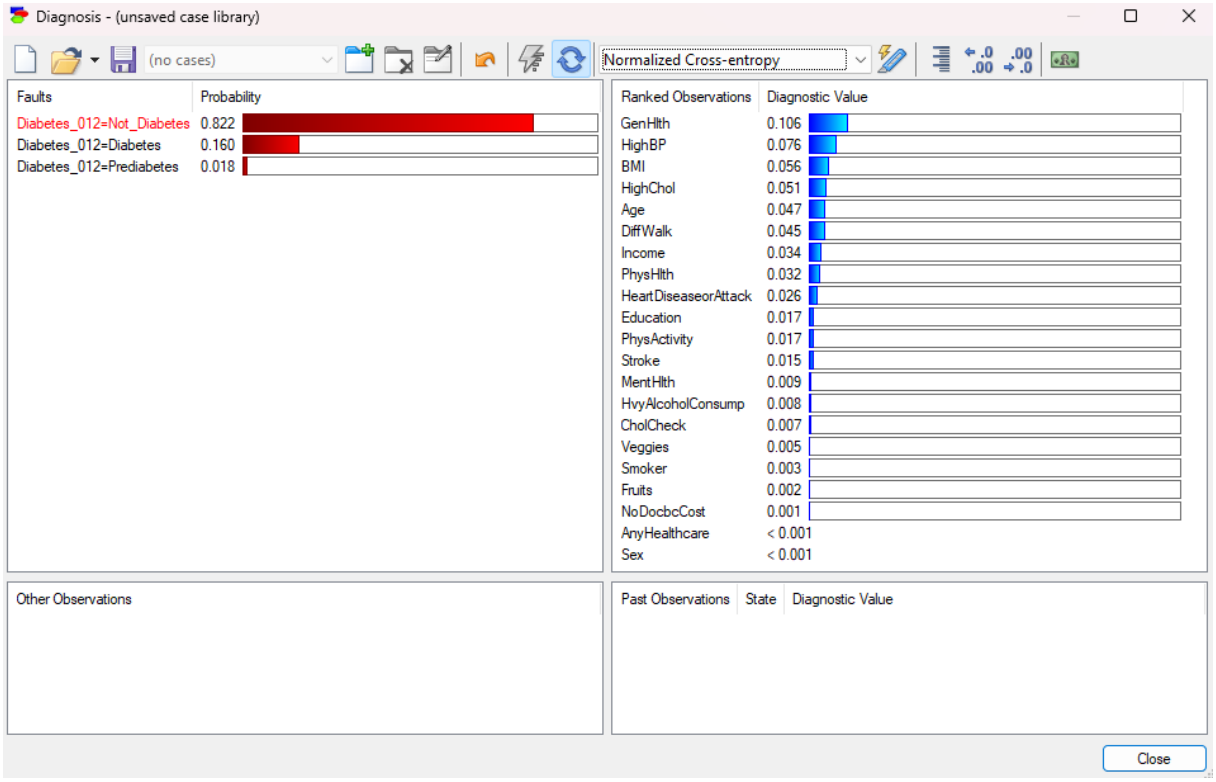
- **Diabetes_012 – BMI** (Cukrzyca – Wskaźnik masy ciała), gdzie cukrzyca jest silnie powiązana z BMI.

- **Diabetes_012 – HighChol** (Cukrzyca – Wysoki cholesterol), gdzie cukrzyca jest powiązana z wysokim cholesterolem.

- **Diabetes_012 – HighBP** (Cukrzyca – Wysokie ciśnienie krwi), gdzie cukrzyca jest powiązana z wysokim ciśnieniem krwi.

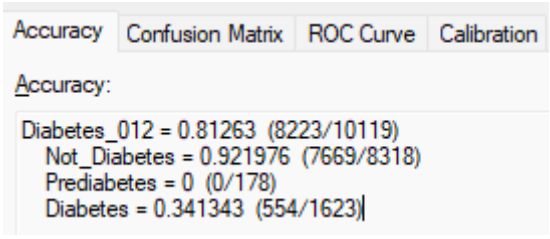
- **Diabetes_012 – AnyHealthcare** (Cukrzyca – Ubezpieczenie zdrowotne), co sugeruje, że osoby z cukrzycą częściej mają ubezpieczenie zdrowotne.

- **Diabetes_012 – DiffWalk** (Cukrzyca – Problemy z poruszaniem), gdzie cukrzyca zwiększa prawdopodobieństwo trudności w poruszaniu się.



*Rysunek 40 Lista zmiennych modelu **Augmented Naive Bayes**, które mają najwyższą wartość diagnostyczną.*

Powyżej znajduje się lista zmiennych modelu **Tree Augmented Naive Bayes**, które mają najwyższą wartość diagnostyczną w celu różnicowania zmiennej, która reprezentuje klasę *Diabetes_012* (cukrzyca). Najwyższą wartość diagnostyczną uzyskała zmienna *GenHlth* (ocena samodzielna stanu zdrowia) uzyskując wynik 0.106. Kolejnymi zmiennymi o wysokiej wartości diagnostycznej są: *HighBP* (wysokie ciśnienie krwi) wynik: 0.076, *BMI* wynik 0.056, *HighChol* (wysoki cholesterol) z wynikiem 0.051, *Age* (kategoria wiekowa) wynik 0.045, *DifWalk* (problemy z poruszaniem się) wynik 0.045. Reszta zmiennych ma mniejszą wartość diagnostyczną od wymienionych wyżej zmiennych. Najmniejszą wartość ma *Sex* (płeć) oraz *AnyHealthcare* (jakiekolwiek ubezpieczenie zdrowotne) z wynikiem mniejszym niż 0.001.



*Rysunek 41 Lista Dokładności dla modelu **Augmented Naive Bayes**.*

Jakość modelu **Augmented Naive Bayes** wynosi ~0.812 czyli około 81,2%. Został on wyliczony na podstawie ilości poprawnie przewidzianych odpowiedzi czy osoba ma cukrzyce, czy jest w tanie przecukrzycowym lub nie ma cukrzycy. Model odgadł 8223 na 10119 przypadków. Jest to wynik lepszy niż uzyskał model naiwnego Bayesa. Przy sprawdzaniu jakości modelu zastosowano metodę **Leave one out**.

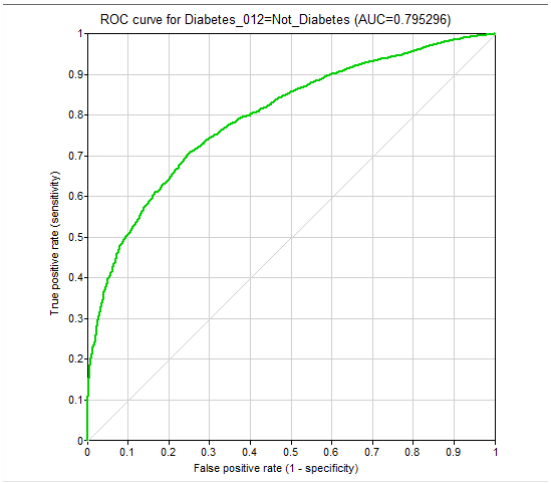
Accuracy | Confusion Matrix | ROC Curve | Calibration

Class node: Diabetes_012

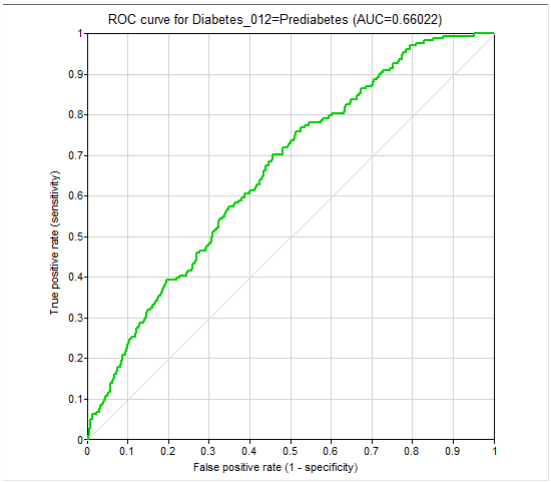
		Predicted		
		Not_Diabetes	Prediabetes	Diabetes
Actual	Not_Diabetes	**7669**	7	642
	Prediabetes	133	**0**	45
	Diabetes	1062	7	**554**

*Rysunek 42 Macierz pomyłek modelu **Augmented Naive Bayes**.D*

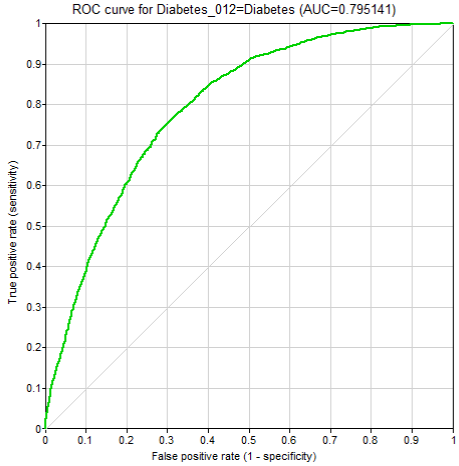
Z macierzy pomyłek dla modelu **Augmented Naive Bayes** wynika, że dla wartości "*Not_Diabetes*" zostało poprawnie przewidzianych 7669 odpowiedzi, a błędnie 649. Dla odpowiedzi "*Prediabetes*" zostało poprawnie przewidzianych 0 odpowiedzi, a błędnie 178. Dla odpowiedzi "*Diabetes*" zostało poprawnie przewidzianych 554 odpowiedzi, a błędnie 1069. Łącznie model przewidział poprawnie 8228 wartości, a błędnie 1896 odpowiedzi.



*Rysunek 43 Krzywa ROC dla modelu **Augmented Naive Bayes** dla odpowiedzi **Nie_Cukrzyk**.*



*Rysunek 44 Krzywa ROC dla modelu **Augmented Naive Bayes** dla odpowiedzi **Stan_przed_Cukrzycowy**.*



*Rysunek 45 Krzywa ROC dla modelu **Augmented Naive Bayes** dla odpowiedzi **Cukrzyk**.*

4. Porównanie wyników jakości modeli na podstawie AUC

Wśród badanych modeli najlepszy okazał się model ***Naive Bayes***, który najlepiej sobie poradził w przypadku Stan_przedcukrzycowy, dla którego AUC (pole pod wykresem krzywej ROC) wyniosło ~0,699, Cukrzyk AUC wyniosło ~0,796, a dla stanu *Nie_cukrzyk* AUC to 0,796. W przypadku stanów *Stan_przedcukrzycowy* oraz *Cukrzyk* problem niewysokiej skuteczności wynikał z małej ilości takich przypadków w naszym zbiorze danych (75 oraz ~1500 rekordów).

Bardzo podobne AUC dla stanów *Nie_cukrzyk* oraz *Cukrzyk* uzyskały modele **Tree Augmented Naive Bayes** oraz **Augmented Naive Bayes**, jednak dla *Stan_przedcukrzycowy ich* AUC wyniosły kolejno ~0.671 oraz ~0,660.

Model **Bayesian Search** okazał się najsłabszy i miał najmniejsze pola pod wykresem krzywej:
- *Stan_przedcukrzycowy* AUC = ~0,5427,
- *Cukrzyk* AUC = ~0,615
- *Nie_cukrzyk* AUC = ~0,616.

Model **PC** poradził sobie wyraźnie lepiej od wcześniej opisanego **Bayesian Search,** ponieważ jego AUC dla *Stan_przedcukrzycowy wyniosło ~0,652,* dla *Cukrzyk ~0,6915,* natomiast dla *Nie_cukrzyk ~0,6921.*

5. Wnioski

Udało się stworzyć model **naiwnego Bayesa**, Model **TAN** - Tree Augmented Naive Bayes, Model **ANB** - Augmented Naive Bayes, model **Bayesian Search** oraz model **PC**. Niestety przy próbie uruchomienia diagnozy dla modelu **PC** program GeNIe Academic 4.1 przestał działać co spowodowało utracenie niezapisanych postępów pracy. Przy ustawieniu properties dla *Age* orz *GenHlt* program wyrzucał błąd „Out of memory". Proces przeprowadzono kilkukrotnie bez sukcesu. Sama walidacja za pomocą metody „Leave one out" dla modelu PC zajęła około 25 minut.

W projekcie została wykorzystana baza danych osób chorch na cukrzyce, w stanie przedcukrzycowym oraz zdrowych. Ankiete przeprowadzono w Ameryce w roku 2015. Bazę ograniczono do 10119 rekordów. Wszystkie modele poddano analizie i obserwacji. Sprawdzono między innymi **listę krawędzi modeli z najsilniejszą mocą wpływu**, **listę zmiennych modeli, które mają najwyższą wartość diagnostyczną** w celu zróżnicowania zmiennej, która reprezentuje klasę Diabietes_012 (stan cukrzycy lub jej braku) wraz z interpretacją. Następnie sprawdzono **jakość modeli**, **czułość** i **specyficzność**, **macierz pomyłek** oraz wygenerowano **wykresy krzywych ROC**. Z analizowanych modeli najlepszym pod względem jakości okazał się być model Bayesian Search, który uzyskał wynik 82,2%. Modele ANB i TAN są bardzo podobne pod względem wyników. Ich jakość to odpowiednio 81,2% i 80,8%. Najsłabiej jakościowo wypadł model naiwnego Bayesa uzyskując wynik 77,9%. **Na podstawie wyników z tabel *Accuracy*** wydawałoby się, że najlepszym modelem i najbardziej dokładnym jakościowo jest model **Bayesian Search,** ale jest to błędna analiza ponieważ model ten zadziałał poprawnie tylko dla stanu *Nie_cukrzyk*. Jego słaba klasyfikacja prawdopodobnie wynika z specyfiki zbioru danych, na których pracowaliśmy, gdzie ponad 82% rekordów to były przypadki osób zdrowych, natomiast *Cukrzycy stanowią* około *17%* i niecały *1%* to rekordy z osobami z *Stan_przed_cukrzycowy.* ***Z porównania modeli*** *na podstawie pola pod wykresem krzywej ROC (***AUC***) wynika, że najlepszy okazał się* model ***Naive Bayes****, który najlepiej sobie poradził z dwoma, problematycznymi stanami, wymienionymi wyżej. Modele* **Tree Augmented Naive Bayes** oraz **Augmented Naive Bayes** poradziły nawet marginalnie lepiej w przypadku *Nie_Cukrzyk* oraz *Cukrzyk*, ale zauważalnie gorzej w przypadku *Stan_przed_cukrzycowy.* Najgorzej w tym porównaniu wypadł właśnie **Bayesian Search**, natomiast mimo trudności w pracy z modelem **PC**, model ten plasuje się po środku skuteczności między modelami najlepszymi i najsłabszym.