POLITECHNIKA WROCŁAWSKA
WYDZIAŁ Informatyki i Telekomunikacji

Algorytmy wspomagania decyzji

Wykorzystanie sieci Bayesowskich w diagnozowaniu cukrzycy

Kierunek: Informatyczne Systemy Automatyki

Prowadzący:
Mgr inż. Cyprian Mataczyński

Autorzy:
Jakub Mazur 247379
Przemysław Marciniak 247331

Wrocław 2023

Motywacja:

Tym co stanowiło motywację do podjęcia wybranego tematu, było z jednej strony zainteresowanie tematami około medycznymi a z drugiej to, że obecnie czas oczekiwania na przyjęcie do szpitala lub wizytę u specjalisty jest bardzo długi. Zastosowanie algorytmów wspomagania decyzji do wstępnej klasyfikacji pacjentów na tych którzy prawdopodobnie nie wymagają leczenia oraz na tych dla których konsultacja lekarska jest wysoce wskazana, może pozwolić zredukować wielkość tego problemu.

Cel projektu:

Celem projektu jest przygotowanie sieci Bayesowskiej diagnozującej pacjentów pod kątem występowania cukrzycy. Program na podstawie dostarczonych wyników badań pacjenta miałby oceniać ryzyko wystąpienia choroby i w razie potrzeby zasugerować skierowanie pacjenta na dalsze badania.

Założenia projektowe:

Do realizacji projektu wykorzystany został zbiór danych „Pima Indians Diabetes Database" udostępniony na stronie internetowej https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Ze względu na to, że w zbiorze występują tylko kobiety między 21 a 81 rokiem życia, uzyskane w trakcie projektu wyniki będą adekwatne tylko względem tej grupy. Poszczególne kolumny zostały podzielone na podzbiory o ustalonych przedziałach wartości. Aby to osiągnąć zastosowano dwa podejścia do rozwiązania tego problemu. Pierwsze z nich bazuje na tradycyjnych metodach analizy i eksploracji danych, natomiast drugie zakłada przygotowanie danych z wykorzystaniem opinii eksperta (lekarza medycyny), więcej szczegółów w rozdziale **Program**.

Do przygotowania sieci Bayesowskiej został wykorzystany język Python oraz następujące biblioteki:

- pgmpy
- bnlearn
- pandas
- numpy
- matplotlib
- seaborn
- sklearn

Baza danych:

Baza danych (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database) jest oparta o dużo większą bazę, z której zostały wyekstrahowane poprawnie zebrane dane pacjentek. Baza danych zawiera dane personalne oraz wyniki badań 768 pacjentek, z czego u 268 wykryto cukrzyce. Zbiór zawiera 9 kolumn:

- Pregnancies – liczba przebytych ciąż (zakres wartości od 0 do 17),
- Glucose – stężenie glukozy w osoczu po 2 godzinach w doustnym teście tolerancji glukozy (zakres wartości od 0 do 199),
- BloodPressure – rozkurczowe ciśnienie krwi (zakres wartości od 0 do 122 mm Hg),
- SkinThickness – grubość fałdu skórnego tricepsa (zakres wartości od 0 do 99 mm),
- Insulin – 2-godzinna insulina w surowicy (zakres wartości od 0 do 846 mu U/ml),
- BMI – wskaźnik masy ciała (zakres wartości od 0 do 67,1 kg/m^2),
- DiabetesPedigreeFunction (DPF) – Funkcja rodowodu cukrzycy, oblicza prawdopodobieństwo wystąpienia cukrzycy w zależności od wieku pacjenta i historii jego rodziny z cukrzycą (zakres wartości od 0,08 do 2,42),

- Age – wieka (zakres wartości od 21 do 81),
- Outcome – zmienna binarna, 0 - osoba zdrowa, 1 - osoba chora na cukrzyce.

Niektóre kolumny zawierały błędne dane (występowały wartości z logicznego punktu widzenia niemożliwe). Dla tych pacjentek dokonano uzupełnienia brakujących danych z wykorzystaniem dwóch, różnych metod. Pierwszą z nich było wykorzystanie regresji Bayesowskiej, natomiast drugą było wykorzystanie średniej arytmetycznej. Kolumny, które wymagały korekty to:

- Glucose – niemożliwe jest, aby glukoza wynosiła 0 (błąd wśród 5 pacjentek)
- BloodPressure – niemożliwe jest, aby ciśnienie krwi wynosiło 0 (błąd wśród 35 pacjentek)
- BMI – niemożliwe jest, aby BMI wynosiło 0 (błąd wśród 11 pacjentek)
- SkinThickness – niemożliwe jest, aby grubość fałdu skórnego wynosiła 0 (błąd wśród 227 pacjentek)

Ze względu na bardzo duży brak danych w przypadku kolumny SkinThickess (~29.5%), podjęta została decyzja o usunięciu jej z bazy danych. Dla wszystkich pozostałych kolumn dokonana została estymacja wartości z wykorzystaniem wcześniej wymienionych metod.

Program:

Jak wspomniano w rozdziale **Założenia projektowe** zbiór danych został podzielony na podzbiory w celu uproszczenia procesu uczenia i zmniejszeniu wymagań sprzętowych potrzebnych do wyuczenia modelu. Pierwsze podejście wykorzystuje metody analizy i eksploracji danych. Na początku tworzony jest histogram w celu obserwacji rozkładu danych. Następnie sąsiednie biny są ze sobą łączone, tak aby zachować pierwotny kształt wykresu. Proces jest powtarzany do momentu, w którym nie będzie możliwe dokonanie kolejnych połączeń podzbiorów bez utraty istotnych informacji o rozkładzie danych. Kod prezentuje się następująco:

```
1  def get_histogram_step(df:pd.DataFrame, label:str, step:int):
2      fmax = int(df[label].max())
3      fmin = int(df[label].min())
4      lhistogram = []
5      ldomain = []
6      for x in range(fmin,fmax-step,step):
7          temp = (df[label]>=x) & (df[label]<x+step)
8          lhistogram.append(temp.sum())
9          ldomain.append(x)
10     plt.plot(ldomain,lhistogram)
11     plt.bar(ldomain,lhistogram)
```

Rysunek 1 Generowanie histogramu w oparciu o biny (wszystkie biny tej samej szerokości)

```
1  def get_histogram_bins(df:pd.DataFrame, label:str, bins:list):
2      plt.hist(data[label],bins)
```

Rysunek 2 Generowanie histogramu w oparciu o listę binów

Uzupełnienie błędnych wartości w kolumnach Glucose, BMI oraz BloodPressure przy wykorzystaniu regresji bayesowskiej prezentuje się następująco:

```python
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.linear_model import BayesianRidge
import numpy as np
#Defining function for regressing data
def regressBayess(data:pd, Column_to_fix:str):
    #Constant, columns with data used for regression
    data_col_for_model = ['Pregnancies','Insulin','DiabetesPedigreeFunction','Age','Outcome']
    subset_names = data_col_for_model
    #Preparing subset with required data (good data + column with incomplete data)
    subset_names.append(Column_to_fix)
    data_subset = data[subset_names]
    #Logical condition for filtering bad data (based on dataset analysis)
    data_filter = (data_subset[Column_to_fix] != 0)
    #Preparing subset w/o incomplete data #data_subset[Column_to_fix] == 0
    data_for_regresion = data_subset.loc[~data_filter].drop(columns=Column_to_fix)
    #Subset which will be updated with generated data
    data_for_after_regresion = data_subset.loc[~data_filter]
    good_rows_x = data_subset.loc[data_filter].drop(columns=Column_to_fix)
    good_rows_y = data_subset[Column_to_fix].loc[data_filter]

    #Random state for stabilizing results
    X_train, X_test, y_train, y_test = train_test_split(good_rows_x, good_rows_y, test_size = 0.15, random_state = 37)

    #Creating model
    model = BayesianRidge()
    model.fit(X_train, y_train)

    #Predicting data with regression
    prediction = model.predict(X_test)

    #Generating R2 score
    print(f"R2 score :", r2_score(y_test, prediction))
    fixed_data = model.predict(data_for_regresion)
    data_for_after_regresion[Column_to_fix] = fixed_data
    data[Column_to_fix].loc[~data_filter] = fixed_data
```

Rysunek 3 Uzupełnienie rekordów regresją bayesowską

```python
regressBayess(data,'Glucose')
regressBayess(data,'BMI')
regressBayess(data,'BloodPressure')

_SavePath = 'data/healthcare/diabetesBayessianRegressed.csv'
data.to_csv(path_or_buf=_SavePath, sep=';')
```

Rysunek 4 Zapisanie zmian do pliku

Poniżej przedstawiono przyjęte przedziały, należy zwrócić uwagę że wartości wpisane w nawiasach są otwartą granicą binu poprzedniego oraz zamkniętą granicą binu kolejnego (na przykładzie Age, bin pierwszy zawiera wartości z przedziału $\langle 21; 24)$, bin drugi wartości z przedziału $\langle 24; 30)$ itd.).

```python
AgeBins =[24,30,38,45,52,60,67]
BMIBins =[23,27,32,35,39,44,50]
BloodPressureBins = [48,64,76,86,108]
InsulinBins = [1,50, 100, 200, 400]
PregnanciesBins =[1,3,6,9,12]
GlucoseBins = [70,90,106,122,136,152,170,188]
DPFBins = [15,40,60,75,100,140]
```

Rysunek 5 Przedziały binów wyznaczone na podstawie histogramów

Drugie podejście użyte do wyznaczenia podzbiorów zakłada wykorzystanie opinii eksperta. W tym celu skontaktowano się z lekarzem medycyny, który przypisał każdej kolumnie najistotniejsze (z punktu widzenia medycyny) przedziały wartości. Podział wygląda następująco:

- BMI:
    1. $x < 18{,}5$
    2. $18{,}5 \leq x < 25$
    3. $25 \leq x < 30$
    4. $x \geq 30$
- Pregnancies:
    1. $x = 0$
    2. $x = 1$ & $x = 2$
    3. $x \geq 3$

- Glucose:
    1. $x < 140$
    2. $140 \leq x < 200$
    3. $x \geq 200$
- BloodPressure:
    1. $x < 80$
    2. $80 \leq x < 85$
    3. $85 \leq x < 90$
    4. $x \geq 90$
- Insulin:
    1. $x < 16$
    2. $16 \leq x \leq 166$
    3. $x > 166$
- Age:
    1. $x < 56$
    2. $56 \leq x < 70$
    3. $x \geq 70$
- DiabetesPedigreeFunction- w tym wypadku lekarz nie podał przedziałów ze względu na brak wiedzy w tym zakresie. Podział został dokonany tą samą metodą jak w pierwszym podejściu, podział na biny

```python
import pandas as pd
import pgmpy
from pgmpy.models import BayesianNetwork
from pgmpy.estimators import MaximumLikelihoodEstimator
from pgmpy.inference import VariableElimination

#Load dataset
data = pd.read_csv('data/healthcare/diabetesBayessianRegressed.csv',sep=';')
#data = pd.read_csv('data/healthcare/diabetesMean.csv',sep=';')
#_SavePath="data/healthcare/diabetesKnowledgeBinnedBayeReg.csv"
#_SavePath="data/healthcare/diabetesKnowledgeBinnedMean.csv"

#Bayesian network
model = BayesianNetwork([
    ('Pregnancies', 'Outcome'),
    ('Glucose', 'Outcome'),
    ('BloodPressure', 'Outcome'),
    ('Insulin', 'Outcome'),
    ('BMI', 'Outcome'),
    ('DiabetesPedigreeFunction', 'Outcome'),
    ('Age', 'Outcome')
])

ddDataRange = {'Age':{1:'X<56',2:'56<=X<70',3:'X>70'},
               'BMI':{1:'X<18.5',2:'18.5<=X<25',3:'25<=X<30',4:'X>=30'},
               'Pregnancies':{0:'0',1:'1<=X<3',2:'X>=3'},
               'BloodPressure':{1:'X<80',2:'80<=X<85',3:'85<=X<90',4:'X>=90'},
               'Insulin':{1:'X<16',2:'16<=X<166',3:'X=>166'},
               'Glucose':{1:'X<140',2:'140<=X<200'},
               'DiabetesPedigreeFunction':{1:'<0.2',2:'0.2 <= <0.45',3:'0.45 <= <0.8',4:'0.8 <= <1.2',5:'=> 1.2'}
               }
```

Rysunek 6 Przygotowanie binów w oparciu o podejście lekarskie

```python
#Podejście zaproponowane przez specjalistę (lekarz pediatra)
data.loc[data['Age']<56,('Age')] = 1#30
data.loc[(data['Age']>55) & (data['Age']<70),('Age')] = 2#65
data.loc[data['Age']>69,('Age')] = 3#75

data.loc[data['BMI']<18.5,('BMI')] = 1
data.loc[(data['BMI']>=18.5) & (data['BMI']<25),('BMI')] = 2
data.loc[(data['BMI']>=25) & (data['BMI']<30),('BMI')] = 3
data.loc[data['BMI']>=30,('BMI')] = 4

data.loc[(data['Pregnancies']>0) & (data['Pregnancies']<3),('Pregnancies')] = 1
data.loc[data['Pregnancies']>=3,('Pregnancies')] = 2

data.loc[data['Glucose']<140,('Glucose')] = 1
data.loc[(data['Glucose']>=140) & (data['Glucose']<200),('Glucose')] = 2
data.loc[data['Glucose']>=200,('Glucose')] = 3

data.loc[data['BloodPressure']<80,('BloodPressure')] = 1
data.loc[(data['BloodPressure']>=80) & (data['BloodPressure']<85),('BloodPressure')] = 2
data.loc[(data['BloodPressure']>=85) & (data['BloodPressure']<90),('BloodPressure')] = 3
data.loc[data['BloodPressure']>=90,('BloodPressure')] = 4

data.loc[data['Insulin']<16,('Insulin')] = 1
data.loc[(data['Insulin']>=16) & (data['Insulin']<=166),('Insulin')] = 2
data.loc[data['Insulin']>166,('Insulin')] = 3

data.loc[data['DiabetesPedigreeFunction']>1.2,('DiabetesPedigreeFunction')] = 5
data.loc[(data['DiabetesPedigreeFunction']>0.80) & (data['DiabetesPedigreeFunction']<=1.2),('DiabetesPedigreeFunction')] = 4
data.loc[(data['DiabetesPedigreeFunction']>0.45) & (data['DiabetesPedigreeFunction']<=0.80),('DiabetesPedigreeFunction')] = 3
data.loc[(data['DiabetesPedigreeFunction']>0.2) & (data['DiabetesPedigreeFunction']<=0.45),('DiabetesPedigreeFunction')] = 2
data.loc[data['DiabetesPedigreeFunction']<=0.2,('DiabetesPedigreeFunction')] = 1
```

Rysunek 7 Przygotowanie binów w oparciu o podejście lekarskie

Uczenie modelu również przebiegało w dwóch wariantach. Pierwszy model powstał w oparciu o bibliotekę pgmpy, natomiast drugi w oparciu o bnlearn.

```
#Model fit
model.fit(data, estimator=MaximumLikelihoodEstimator)

#Inference
infer = VariableElimination(model)

#Features for prediction
evidence = {'Age':1,'BMI': 2, 'Pregnancies': 0, 'Glucose': 1, 'BloodPressure': 1, 'Insulin': 1}

#Prediction based on features set
diabetes_prob = infer.query(['Outcome'], evidence=evidence)
for x in evidence:
    print(x,":",ddDataRange[x][evidence[x]])
print(diabetes_prob)
```

Rysunek 8 Uczenie modelu- pgmpy

```
#Base dataset
#train = pd.read_csv('data/healthcare/diabetes_train.csv',sep=';')
#test = pd.read_csv('data/healthcare/diabetes_test.csv',sep=';')

#Filled dataset, before binning
#train = pd.read_csv('data/healthcare/diabetesBayessianRegressed.csv',sep=';')
#train = pd.read_csv('data/healthcare/diabetesMean.csv',sep=';')

#Filled dataset, after binning
train = pd.read_csv('data/healthcare/diabetesKnowledgeBinnedBayeReg.csv',sep=';')
#train = pd.read_csv('data/healthcare/diabetesKnowledgeBinnedMean.csv',sep=';')
```

```
#Splitting data
df_data,df_valid = train_test_split(df_train,  test_size=0.2, random_state=0)
df_data
```

```
# Structure learning
DAG = bn.structure_learning.fit(df_data, methodtype='hc', root_node='Outcome', verbose=5)
#Plot DAG -> directed acyclic graph
G = bn.plot(DAG)
#Create model based on earlier structere and parameter learning
model = bn.parameter_learning.fit(DAG, df_data, verbose=5)
```

Rysunek 9 Wczytanie danych, preparacja struktury i modelu z wykorzystanie biblioteki bnlearn

Do oceny skuteczności modeli wykorzystane zostały tabele pomyłek. Dodatkowo wyodrębnione zostały maksymalne i minimalne wartości prawdopodobieństwa dla każdego pola macierzy (True Positive, True Negative, False Positive, False Negative).

```
#Prediction using the Bayesian network model
Pout = bn.predict(model, df=df_valid, variables=['Outcome'])

#Preparing pandas dataframe for determining coonfusion matrix
probablity_table = pd.concat([df_valid['Outcome'].reset_index(drop=True),Pout['Outcome'],Pout['p']],
                             ignore_index=True, axis=1)
probablity_table.columns=['GroundTruth','Prediction','Probability']

#Creating additional column for working comfort (prediction vs ground truth)
probablity_table['Confusion'] = 0
probablity_table['Confusion'].loc[(probablity_table['GroundTruth']==0)&(probablity_table['Prediction']==0)] = "TN"
probablity_table['Confusion'].loc[(probablity_table['GroundTruth']==0)&(probablity_table['Prediction']==1)] = "FP"
probablity_table['Confusion'].loc[(probablity_table['GroundTruth']==1)&(probablity_table['Prediction']==0)] = "FN"
probablity_table['Confusion'].loc[(probablity_table['GroundTruth']==1)&(probablity_table['Prediction']==1)] = "TP"
```

Rysunek 10 Utworzenie macierzy pomyłek

```
#Counting occurences in confusion column
tp=probablity_table['Confusion'].value_counts()["TP"]
fp=probablity_table['Confusion'].value_counts()["FP"]
tn=probablity_table['Confusion'].value_counts()["TN"]
fn=probablity_table['Confusion'].value_counts()["FN"]

#Creating base fo confusion matrix
ConfusionMatrix = [[tp,fn],[fp,tn]]
df_cm = pd.DataFrame(ConfusionMatrix, index = [i for i in ["GroundTrue","GroundFalse"]],
                     columns = [i for i in ["PredictTrue","PredictFalse"]])

#Using seaborn to plot confusion matrix
sns.set(font_scale=1.4)
sns.heatmap(df_cm, annot=True, annot_kws={"size": 16})
plt.show()
```

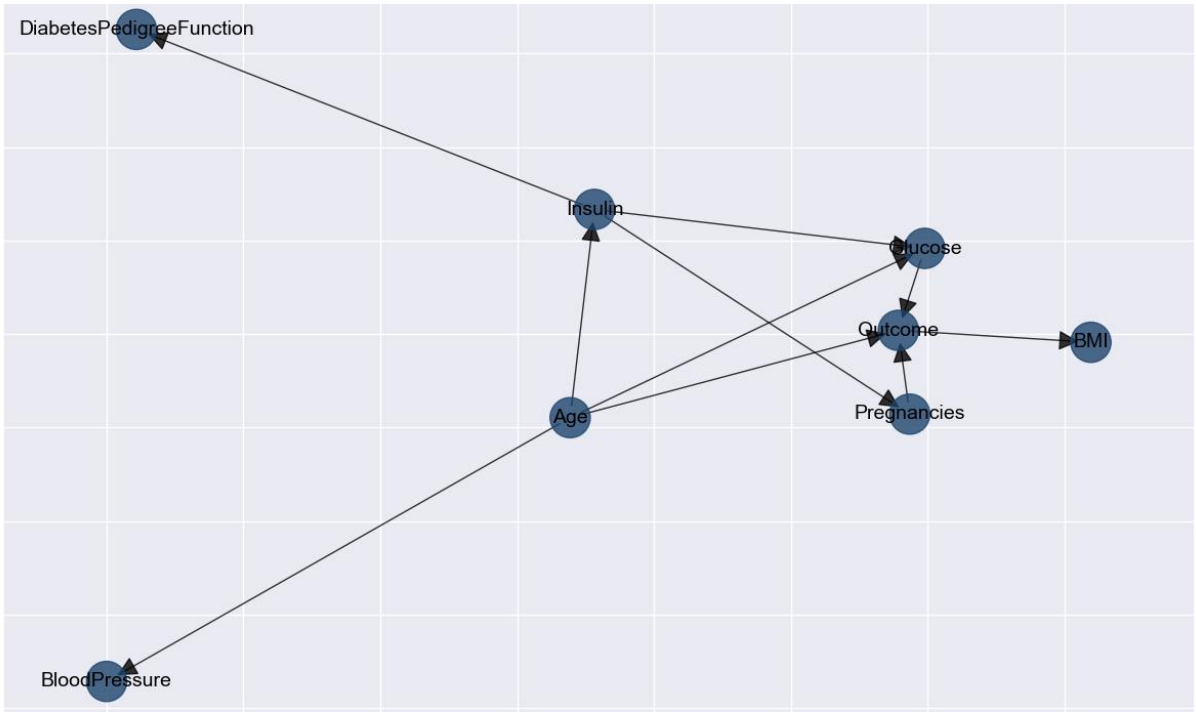
Rysunek 11 Wyświetlenie macierzy pomyłek

```
#Extracting information about min/max probability in every case
probMin=probablity_table['Probability'].loc[probablity_table['Confusion']=='TP'].min()
probMax=probablity_table['Probability'].loc[probablity_table['Confusion']=='TP'].max()
print('TP = {}, Min. Prob. = {}, Max. Prob. = {}'.format(tp,probMin,probMax))

probMin=probablity_table['Probability'].loc[probablity_table['Confusion']=='FP'].min()
probMax=probablity_table['Probability'].loc[probablity_table['Confusion']=='FP'].max()
print('FP = {}, Min. Prob. = {}, Max. Prob. = {}'.format(fp,probMin,probMax))

probMin=probablity_table['Probability'].loc[probablity_table['Confusion']=='TN'].min()
probMax=probablity_table['Probability'].loc[probablity_table['Confusion']=='TN'].max()
print('TN = {}, Min. Prob. = {}, Max. Prob. = {}'.format(tn,probMin,probMax))

probMin=probablity_table['Probability'].loc[probablity_table['Confusion']=='FN'].min()
probMax=probablity_table['Probability'].loc[probablity_table['Confusion']=='FN'].max()
print('FN = {}, Min. Prob. = {}, Max. Prob. = {}'.format(fn,probMin,probMax))
```

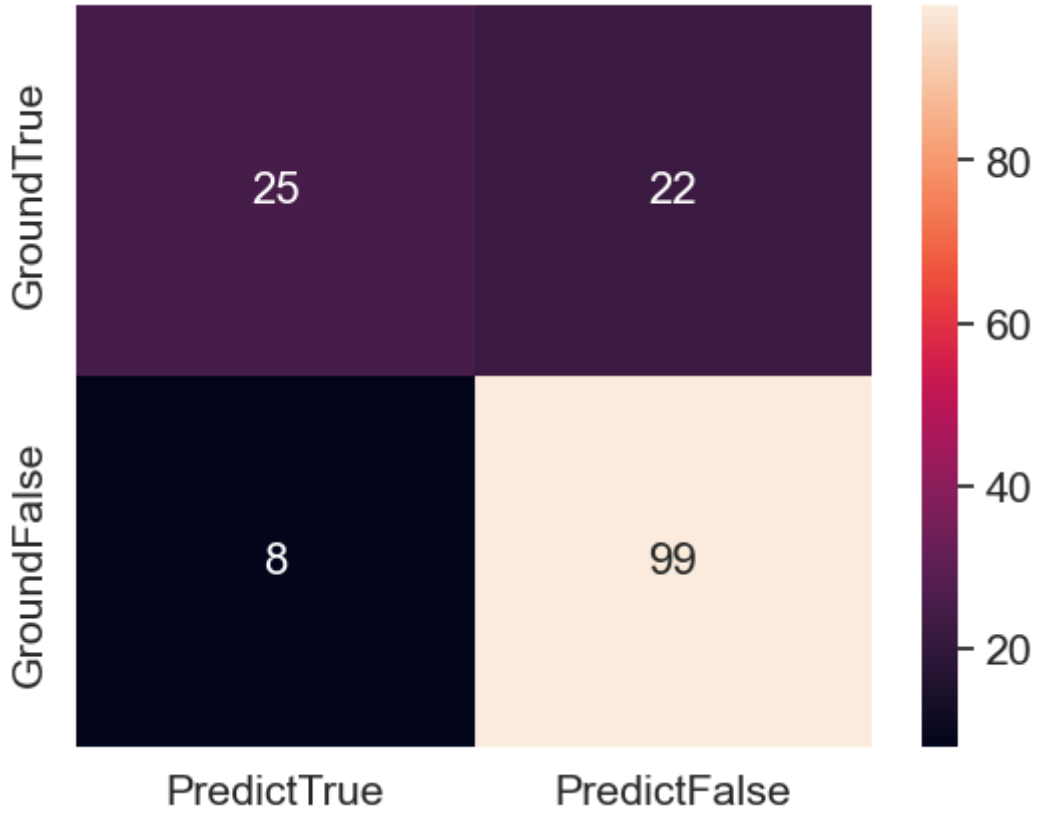
Rysunek 12 Wydobycie informacji o wartościach prawdopodobieństw

Wyniki:

Rezultat działania sieci przedstawiony zostanie przy pomocy grafów. Strzałki wskazują na powiązanie jednego parametru z drugim, natomiast grot strzałki na to, która wielkość wpływa na którą. Im mniejsza długość linii łączącej owe parametry tym silniejsze powiązanie między nimi. Wyniki zostały zaprezentowane dla obydwóch podejść podziału zbiory na podzbiory („Podział lekarski” oraz „Podział histogramy”), na uzupełnienie brakujących wartości (regresja bayesowska oraz średnia arytmetyczna) i na wykorzystany, podczas uczenia, error score (K2, BIC, BDEU).



Rysunek 13 Podział – lekarski, uzupełnienie danych – regresja bayesowska, błąd – K2

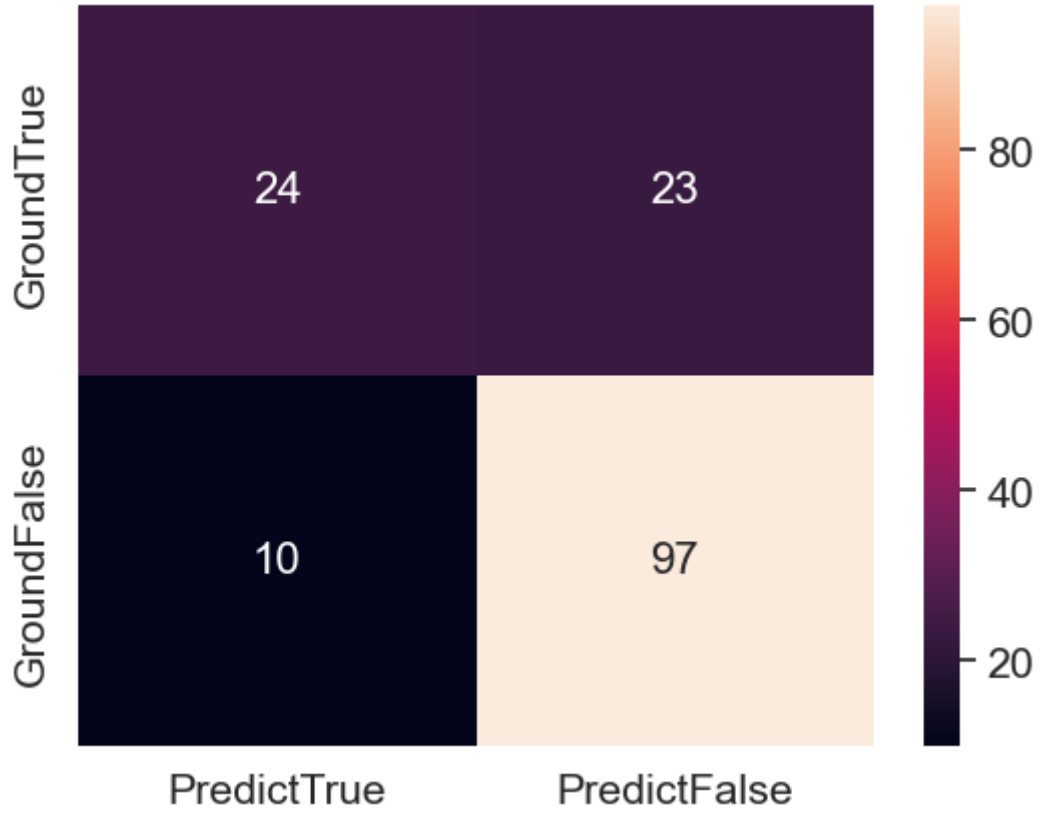


Tablica Pomyłek 1 Podział – lekarski, uzupełnienie danych – regresja bayesowska, błąd – K2

Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 25, Min. Prob. = 0.560962255531582, Max. Prob. = 0.6630779783254445
- FP = 8, Min. Prob. = 0.5609622555315819, Max. Prob. = 0.6630779783254445
- TN = 99, Min. Prob. = 0.5075004286568011, Max. Prob. = 0.8036406775072525
- FN = 22, Min. Prob. = 0.5527869355209691, Max. Prob. = 0.7906153053828006

Rysunek 14 Podział – lekarski, uzupełnienie danych – regresja bayesowska, błąd – BIC

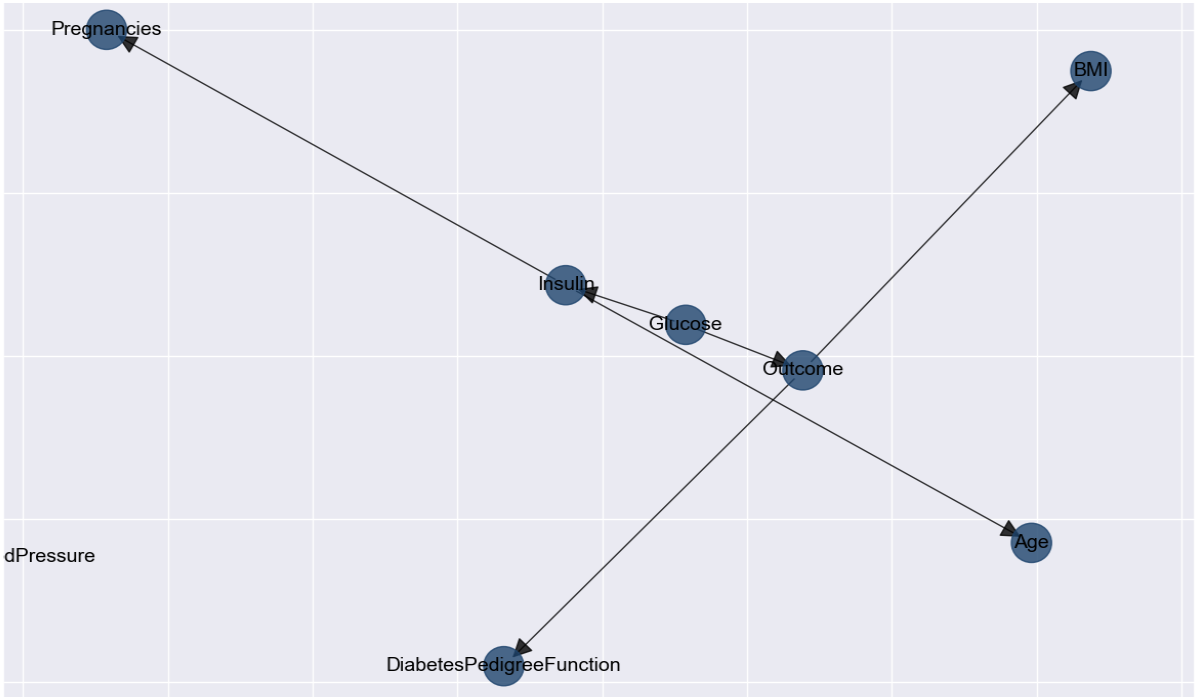


Tablica Pomyłek 2 Podział – lekarski, uzupełnienie danych – regresja bayesowska, błąd – BIC

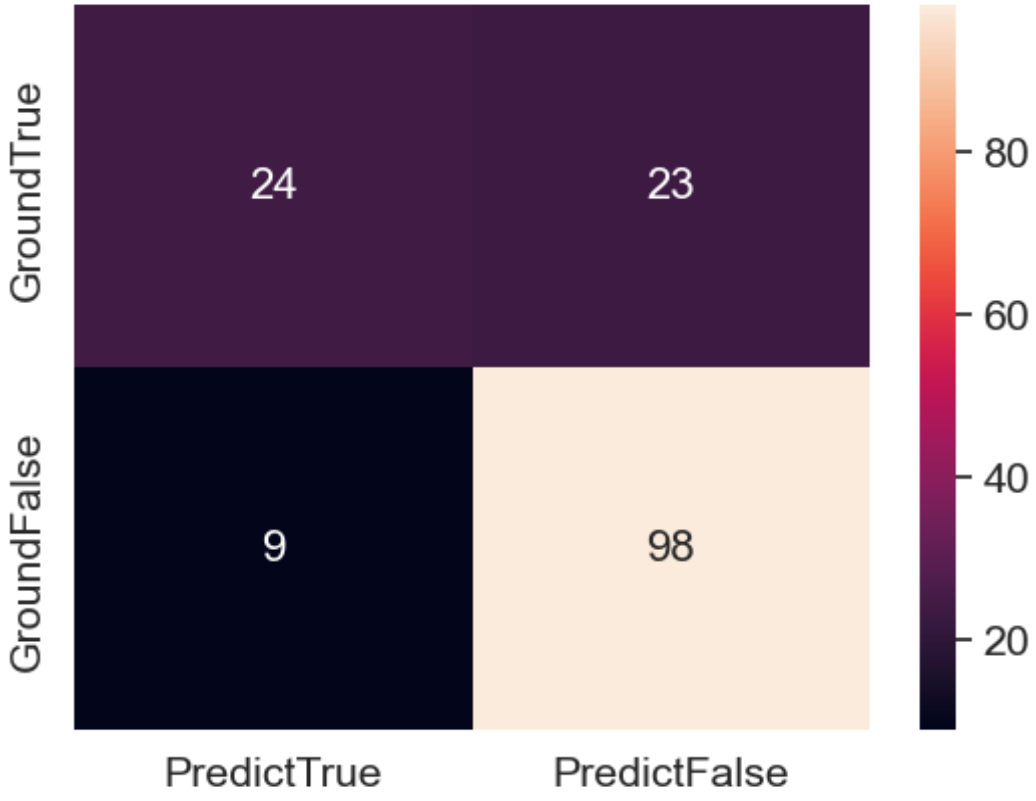
Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 24, Min. Prob. = 0.5076459603915973, Max. Prob. = 0.5758251799870121
- FP = 10, Min. Prob. = 0.5076459603915973, Max. Prob. = 0.5758251799870121
- TN = 97, Min. Prob. = 0.5124927644663098, Max. Prob. = 0.673400126991624
- FN = 23, Min. Prob. = 0.5124927644663098, Max. Prob. = 0.6554394788384637

Parametry porzucone w momencie tworzenia macierzy pomyłek: BloodPressure, DPF.



Rysunek 15 Podział – lekarski, uzupełnienie danych – regresja bayesowska, błąd – BDEU



Tablica Pomyłek 3 Podział – lekarski, uzupełnienie danych – regresja bayesowska, błąd – BDEU

Minimalne i maksymalne wartości prawdopodobieństw:

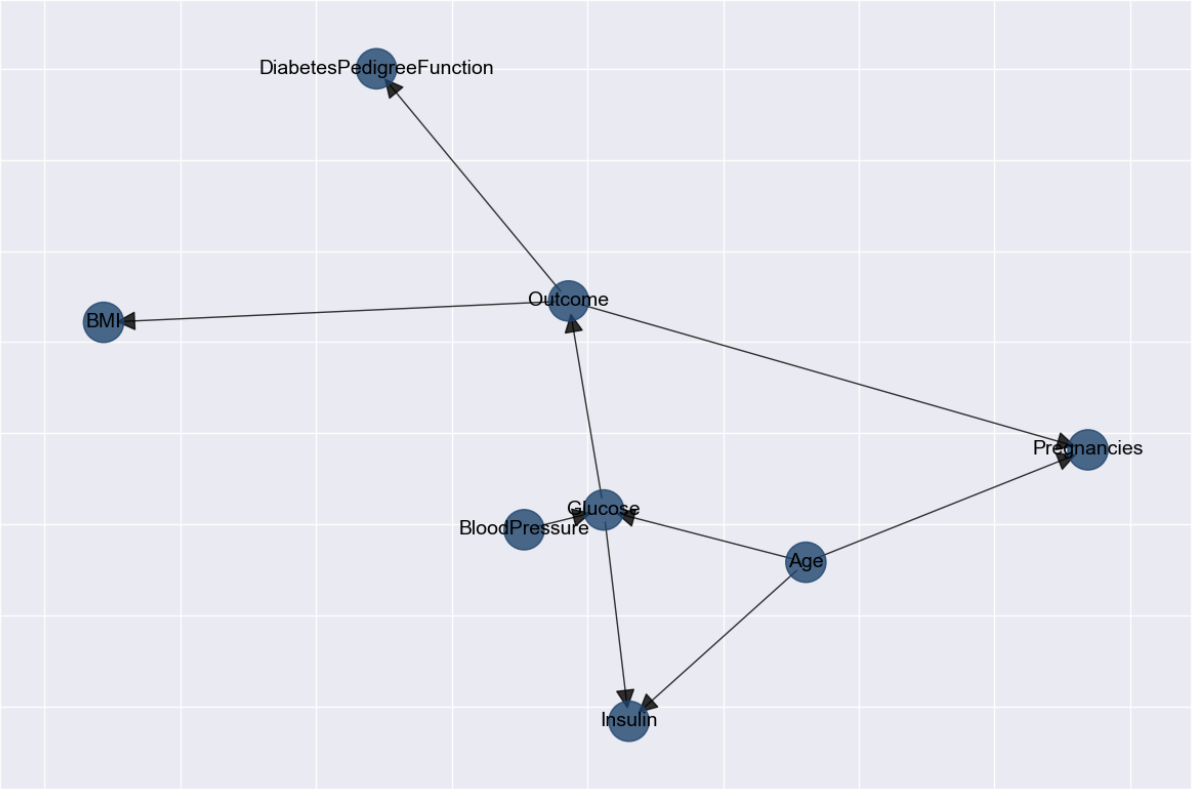
- TP = 24, Min. Prob. = 0.5099902961480683, Max. Prob. = 0.6374830451909156
- FP = 9, Min. Prob. = 0.5099902961480683, Max. Prob. = 0.6374830451909155
- TN = 98, Min. Prob. = 0.5187054735455234, Max. Prob. = 0.7162612878165882
- FN = 23, Min. Prob. = 0.5101489014951446, Max. Prob. = 0.6788472501599758

Parametry porzucone w momencie tworzenia macierzy pomyłek: BloodPressure.

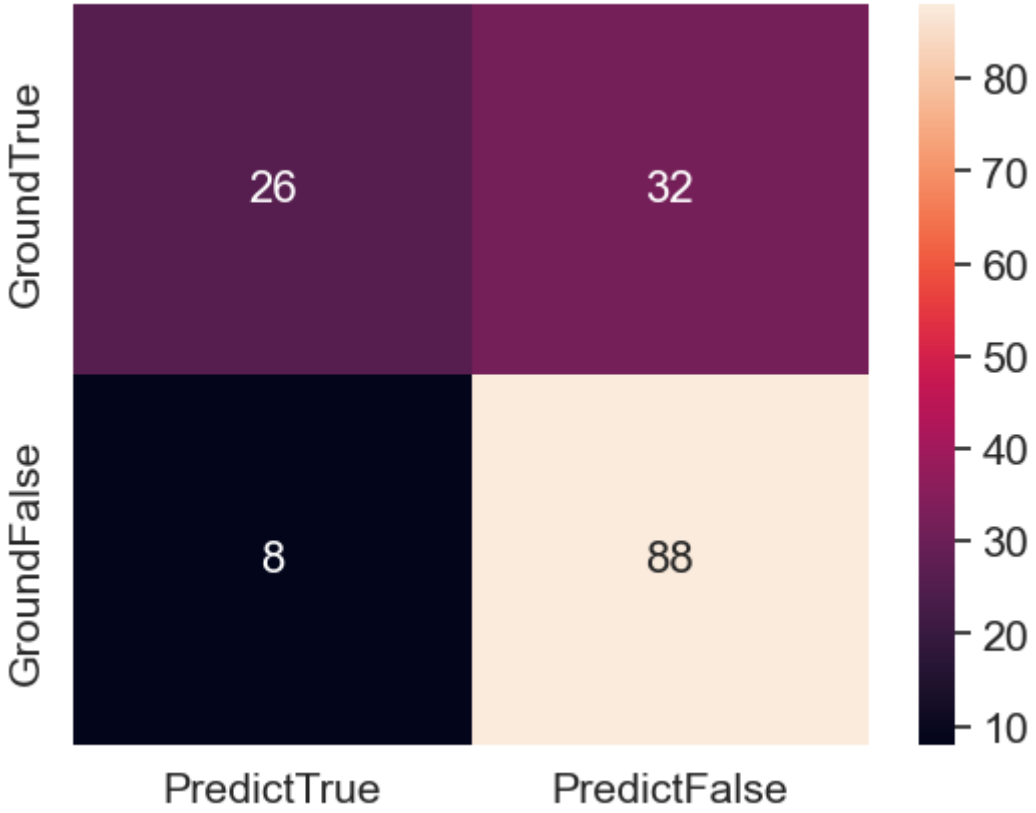
W przypadku błędu K2 wszystkie parametry znajdują się na grafie, a liczba powiązań wynosi 10. Bardzo silne powiązania występują pomiędzy Glucose -> Outcome oraz Pregnancies -> Outcome. Jedynym parametrem, którego wartość jest powiązana z obecnością cukrzycy jest BMI (Outcome -> BMI). Obserwując tablice pomyłek, można zauważyć, że dokładność modelu wynosi 80,519%. Najmniejsze uzyskane prawdopodobieństwo wyniosło 50,750%, natomiast największe 80,364%.

W przypadku błędu BIC na grafie brakuje 2 parametrów: BloodPressure oraz DPF, a liczba powiązań wynosi tylko 5. Najsilniejsze powiązanie występuje pomiędzy Age -> Insulin. Tutaj również jedynym parametrem, którego wartość jest powiązana z obecnością cukrzycy jest BMI. Ze względu na brak powiązań parametrów BloodPressure i DPF, z innymi wielkościami na grafie, te parametry zostały pominięte podczas wyznaczania prawdopodobieństw i obliczania tablicy pomyłek. Dokładność modelu wynosi 78,571%. Najmniejsze uzyskane prawdopodobieństwo wyniosło 50,765%, natomiast największe 67,340%.

W przypadku błędu BDEU wszystkie parametry są widoczne na grafie, ale BloodPressure nie jest powiązane z żadną inną wielkością, dlatego zostanie ona porzucona w dalszym etapie obliczania prawdopodobieństw. Liczba powiązań wynosi 6. Bardzo silne powiązania występują pomiędzy Glucose -> Insulin oraz Glucose -> Outcome. Parametrami, których wartości zależą od obecności cukrzycy są BMI oraz DPF. Dokładność modelu wynosi 79,221%. Minimalne prawdopodobieństwo wynosi 50,999%, a maksymalne 71,626%



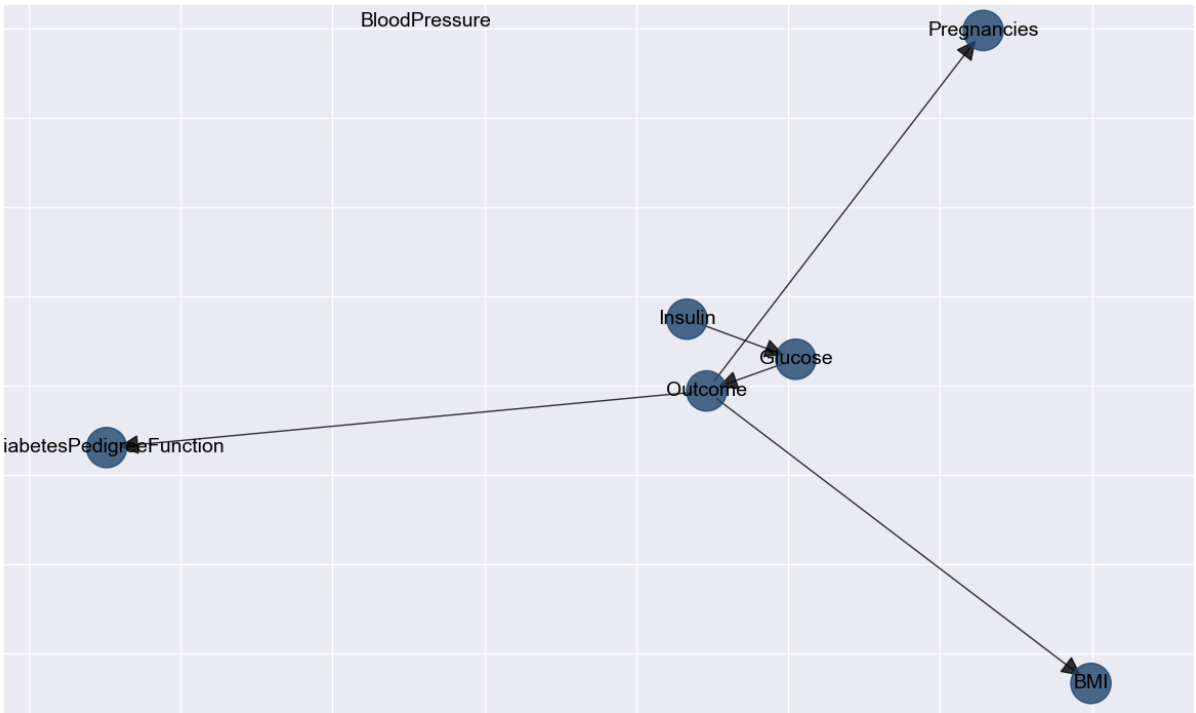
Rysunek 16 Podział – lekarski, uzupełnienie danych – średnia arytmetyczna, błąd – K2



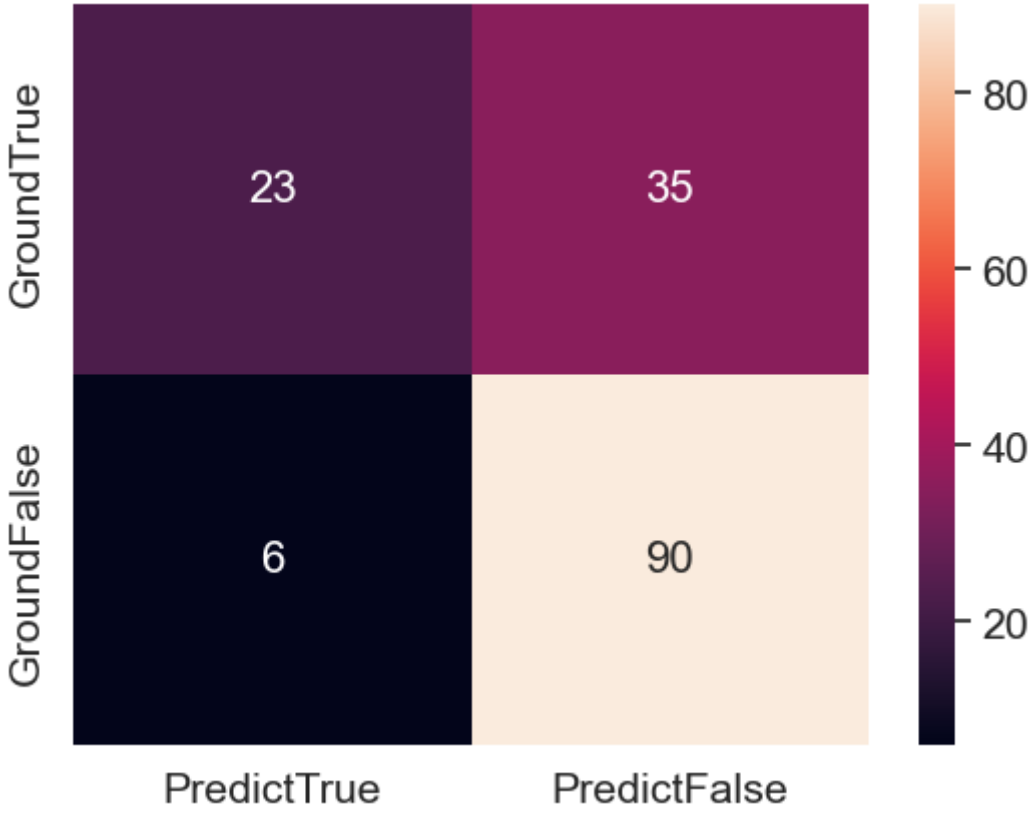
Tablica Pomyłek 4 Podział – lekarski, uzupełnienie danych – średnia arytmetyczna, błąd – K2

Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 26, Min. Prob. = 0.5000739719803626, Max. Prob. = 0.6828083894404463
- FP = 8, Min. Prob. = 0.5000739719803627, Max. Prob. = 0.6828083894404463
- TN = 88, Min. Prob. = 0.5080272725810886, Max. Prob. = 0.7940333125924515
- FN = 32, Min. Prob. = 0.5463155967048513, Max. Prob. = 0.7425297343044079



Rysunek 17 Podział – lekarski, uzupełnienie danych – średnia arytmetyczna, błąd – BIC

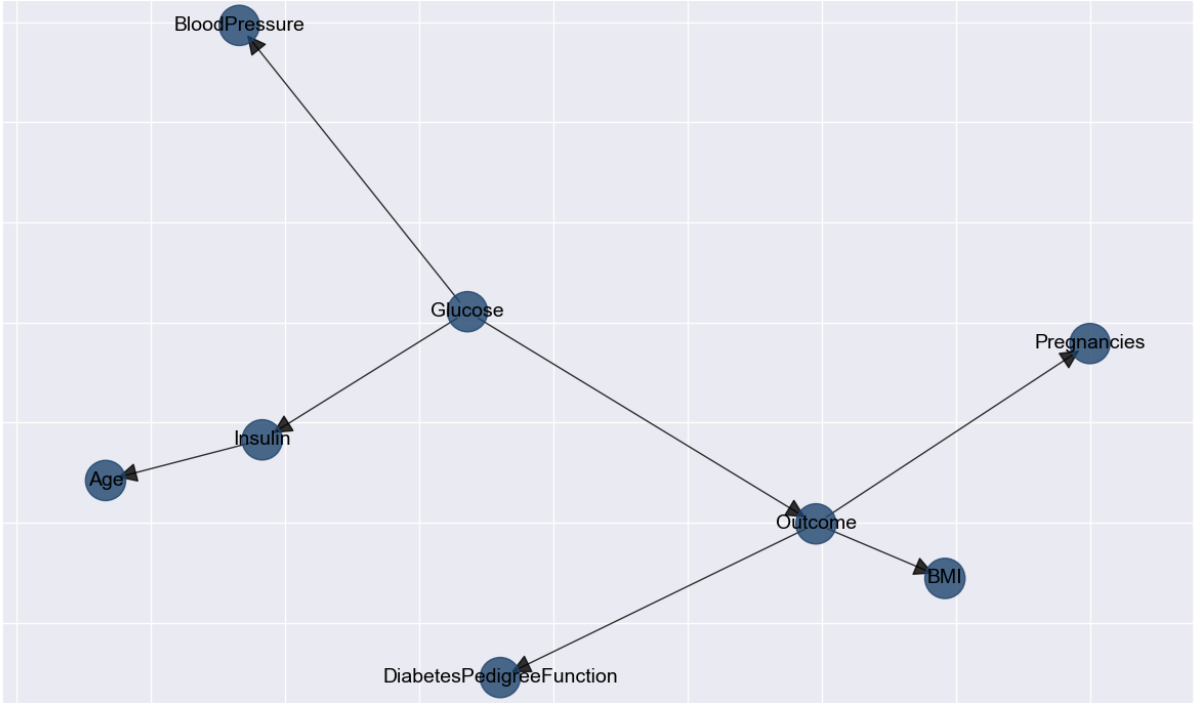


Tablica Pomyłek 5 Podział – lekarski, uzupełnienie danych – średnia arytmetyczna, błąd – BIC

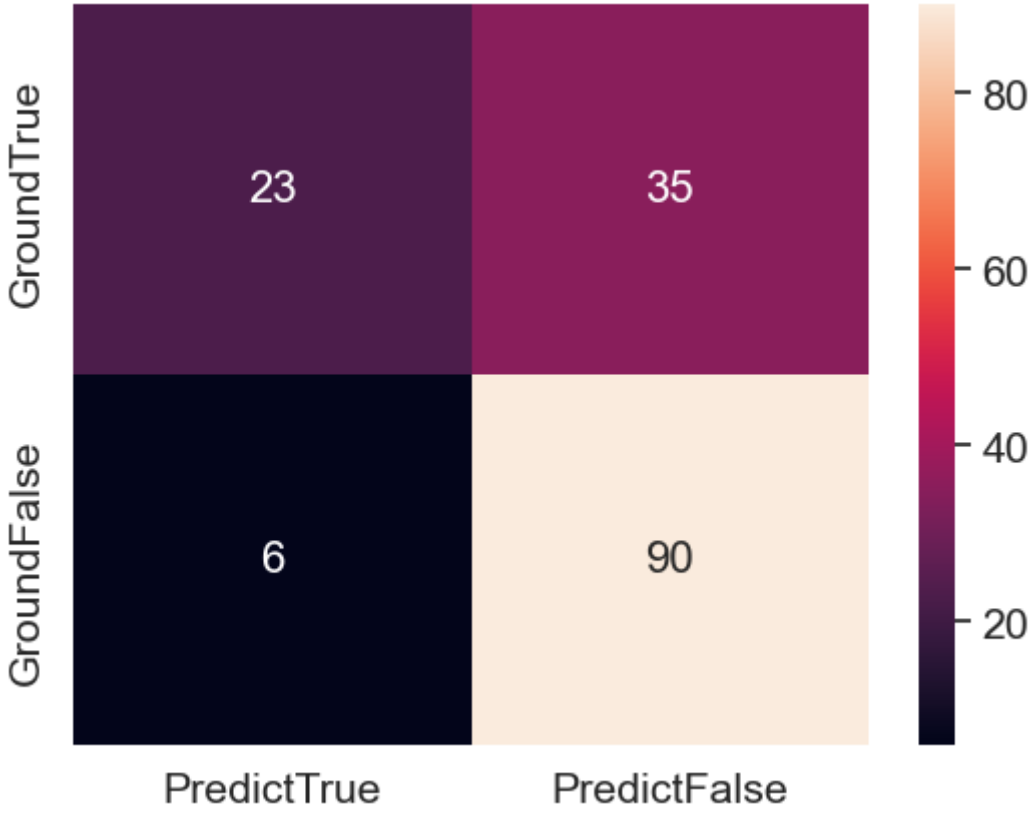
Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 23, Min. Prob. = 0.5453472307566528, Max. Prob. = 0.6566991975326701
- FP = 6, Min. Prob. = 0.5164326796892502, Max. Prob. = 0.6566991975326701
- TN = 90, Min. Prob. = 0.5152815750440917, Max. Prob. = 0.7557749368501574
- FN = 35, Min. Prob. = 0.5061941079795798, Max. Prob. = 0.6983390748419234

Parametry porzucone w momencie tworzenia macierzy pomyłek: BloodPressure, Age.



Rysunek 18 Podział – lekarski, uzupełnienie danych – średnia arytmetyczna, błąd – BDEU



Tablica Pomyłek 6 Podział – lekarski, uzupełnienie danych – średnia arytmetyczna, błąd – BDEU

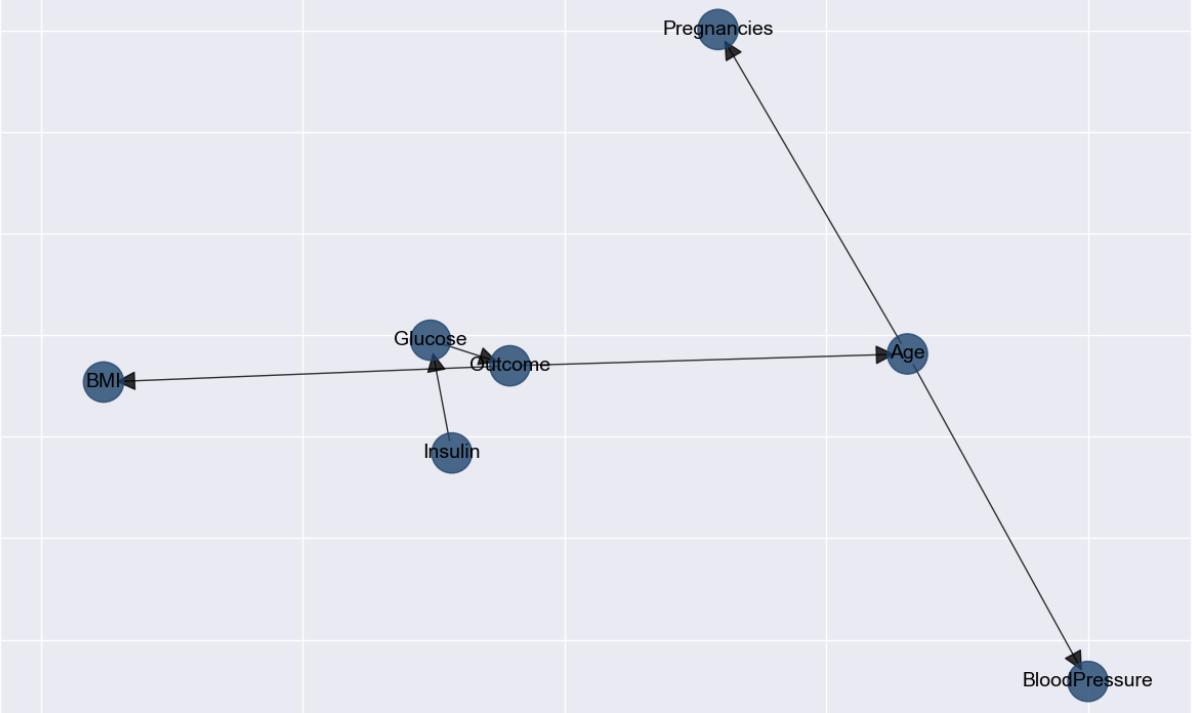
Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 23, Min. Prob. = 0.5453472307566528, Max. Prob. = 0.65669919753267
- FP = 6, Min. Prob. = 0.5164326796892502, Max. Prob. = 0.65669919753267
- TN = 90, Min. Prob. = 0.5152815750440918, Max. Prob. = 0.7557749368501574
- FN = 35, Min. Prob. = 0.5061941079795799, Max. Prob. = 0.6983390748419235

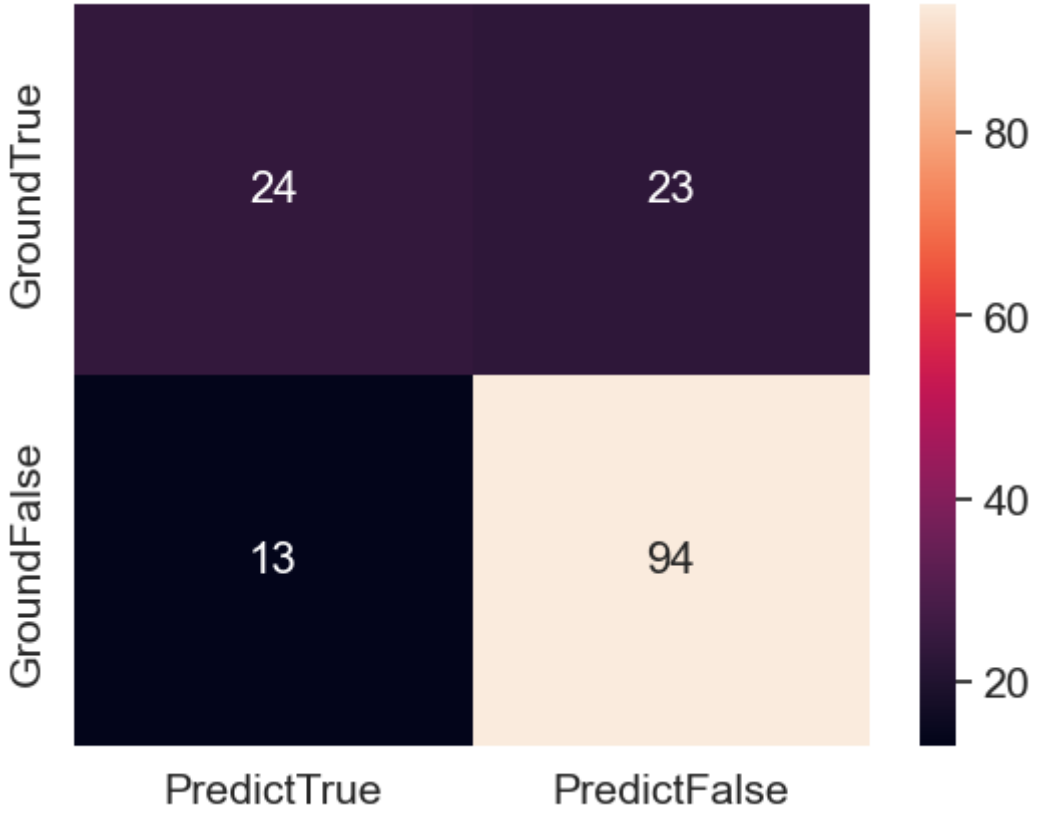
W przypadku błędu K2 wszystkie parametry znajdują się na grafie, a liczba powiązań wynosi 9. Bardzo silne powiązanie występuje pomiędzy BloodPressure -> Glucose. Parametry, których wartości zależą od obecności cukrzycy to BMI oraz DPF. Obserwując tablice pomyłek, można zauważyć, że dokładność modelu wynosi 74,026%. Najmniejsze uzyskane prawdopodobieństwo wyniosło 50,007%, natomiast największe 79,403%.

W przypadku błędu BIC na grafie brakuje jednego parametru- Age, natomiast BloodPressure nie jest powiązany z żadną inną wielkością. Liczba powiązań wynosi 5. Najsilniejsze powiązania występują pomiędzy Insulin -> Glucose oraz Glucose -> Outcome. Parametrami, których wartości są powiązane z obecnością cukrzycy są BMI, Pregnancies oraz DPF. Dokładność modelu wynosi 73,377%. Najmniejsze uzyskane prawdopodobieństwo wyniosło 50,619%, natomiast największe 75,577%.

W przypadku błędu BDEU wszystkie parametry są widoczne na grafie, a liczba powiązań wynosi 7. Silne powiązania występują pomiędzy Outcome -> BMI oraz Insulin -> Age. Dokładność modelu wynosi 73,377%. Minimalne prawdopodobieństwo wynosi 50,619%, a maksymalne 75,577%



Rysunek 19 Podział – histogramy, uzupełnienie danych – regresja bayesowska, błąd – K2

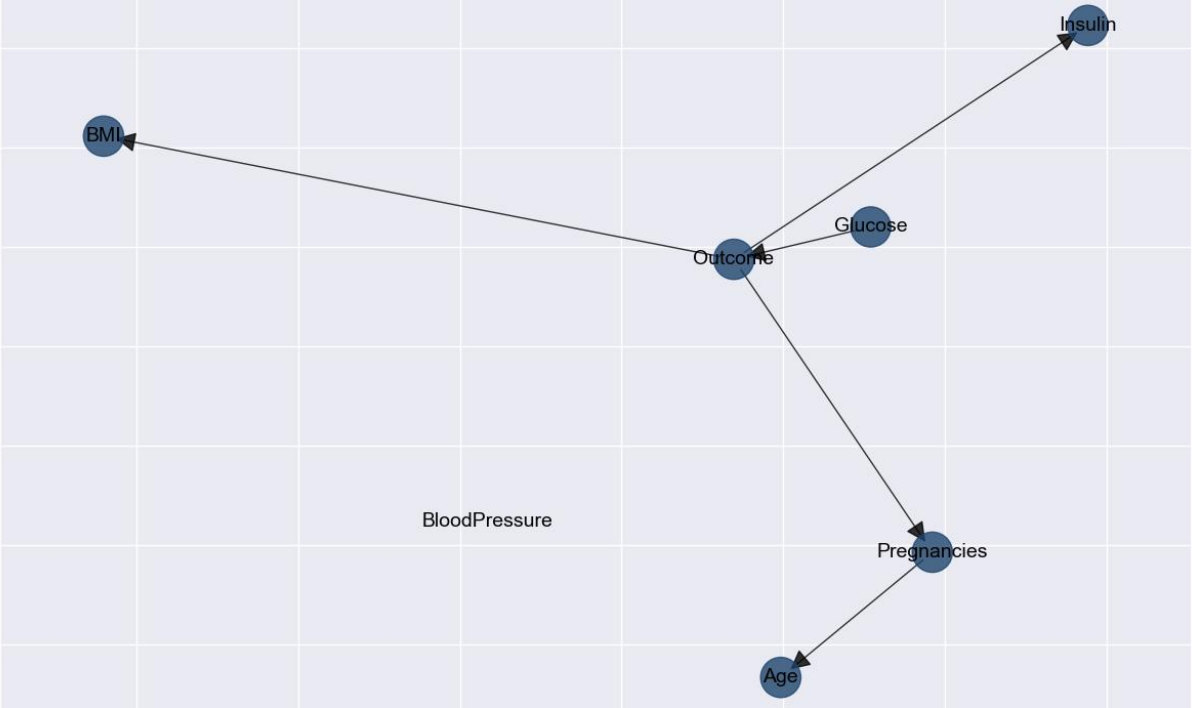


Tablica Pomyłek 7 Podział – histogramy, uzupełnienie danych – regresja bayesowska, błąd – K2

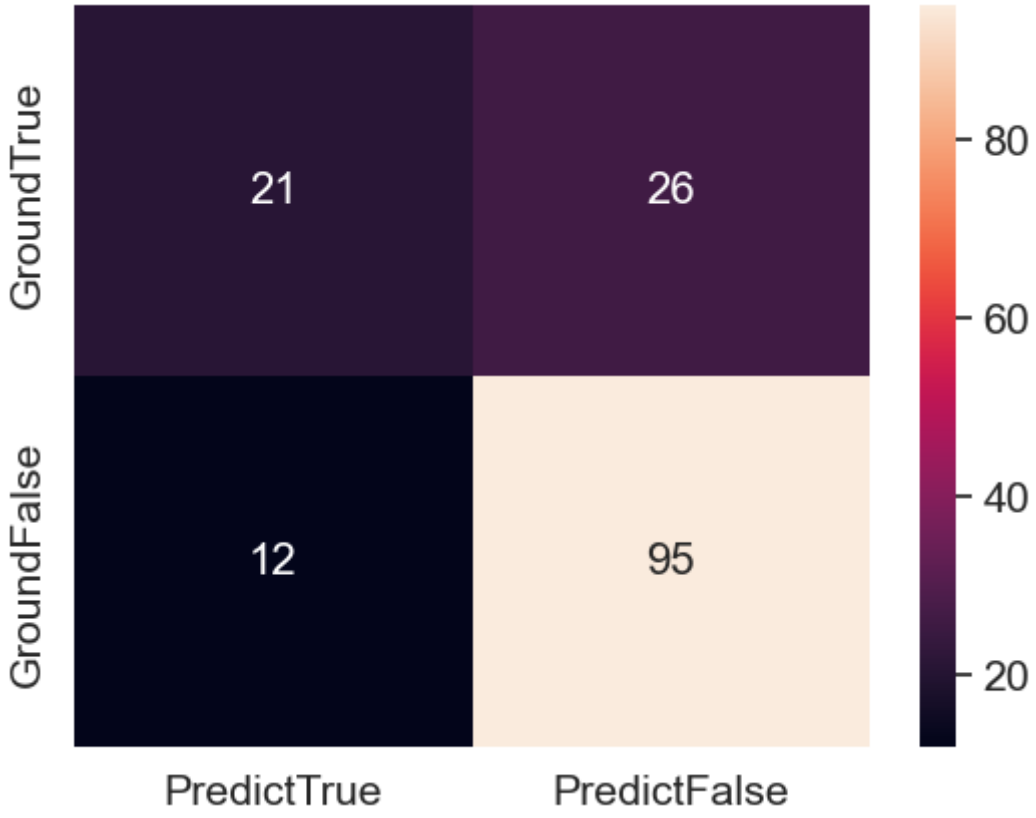
Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 24, Min. Prob. = 0.50313045428628, Max. Prob. = 0.699946605499103
- FP = 13, Min. Prob. = 0.5002503456713857, Max. Prob. = 0.7206761741974547
- TN = 94, Min. Prob. = 0.5136273545676588, Max. Prob. = 0.8716517613893834
- FN = 23, Min. Prob. = 0.5165497613601329, Max. Prob. = 0.7576653713208007

Parametry porzucone w momencie tworzenia macierzy pomyłek: DPF.



Rysunek 20 Podział – histogramy, uzupełnienie danych – regresja bayesowska, błąd – BIC

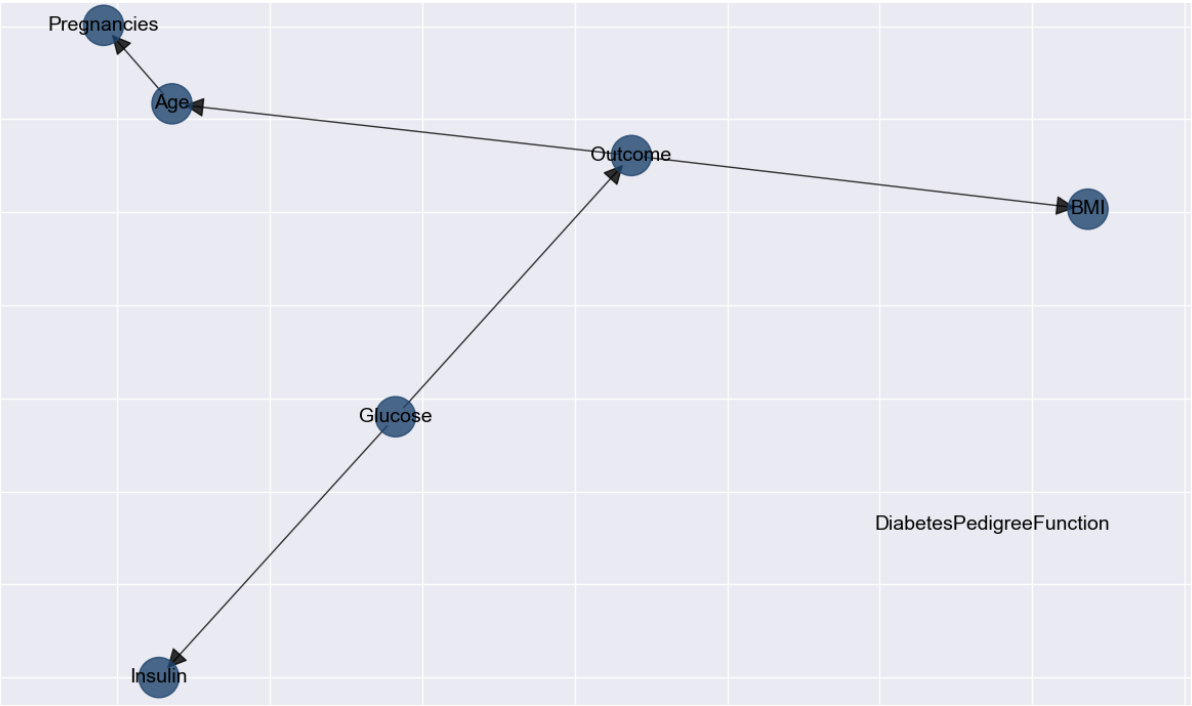


Tablica Pomyłek 8 Podział – histogramy, uzupełnienie danych – regresja bayesowska, błąd – BIC

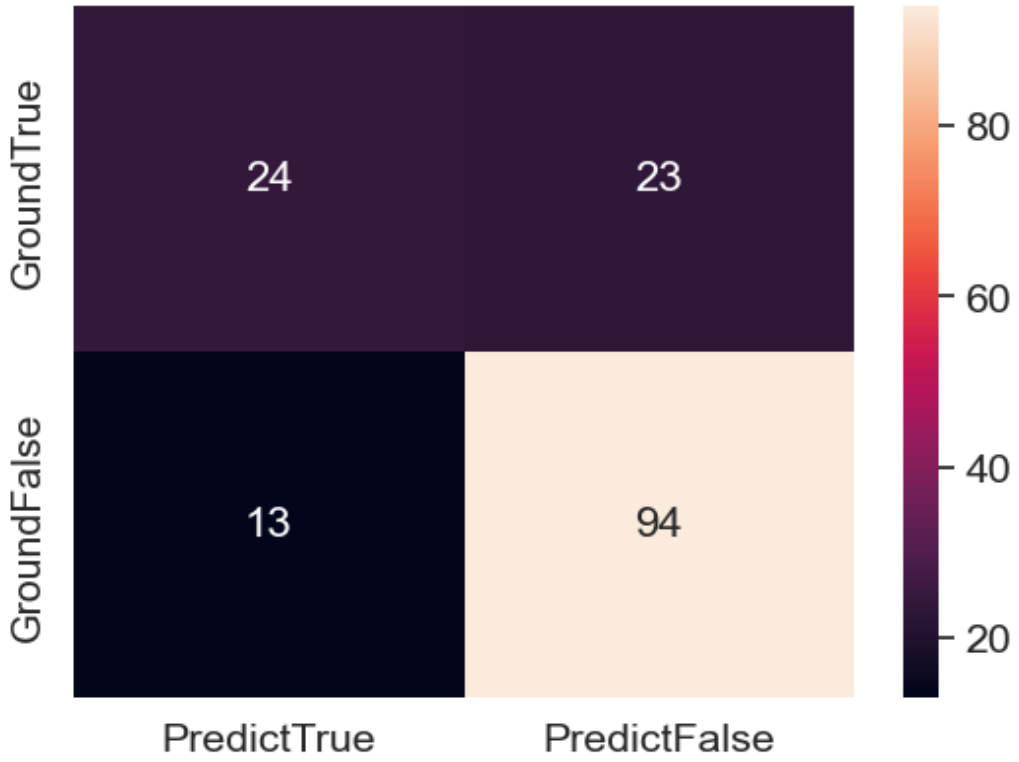
Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 21, Min. Prob. = 0.5089689188194405, Max. Prob. = 0.7190888456774572
- FP = 12, Min. Prob. = 0.5030358510014052, Max. Prob. = 0.6915284692505386
- TN = 95, Min. Prob. = 0.504211409986565, Max. Prob. = 0.855555087547847
- FN = 26, Min. Prob. = 0.5020848701841889, Max. Prob. = 0.6981981267416547

Parametry porzucone w momencie tworzenia macierzy pomyłek: BloodPressure, DPF.



Rysunek 21 Podział – histogramy, uzupełnienie danych – regresja bayesowska, błąd – BDEU



Tablica Pomyłek 9 Podział – histogramy, uzupełnienie danych – regresja bayesowska, błąd – BDEU

Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 24, Min. Prob. = 0.50313045428628, Max. Prob. = 0.699946605499103
- FP = 13, Min. Prob. = 0.5002503456713858, Max. Prob. = 0.7206761741974547
- TN = 94, Min. Prob. = 0.5136273545676588, Max. Prob. = 0.8716517613893834
- FN = 23, Min. Prob. = 0.5165497613601328, Max. Prob. = 0.7576653713208007

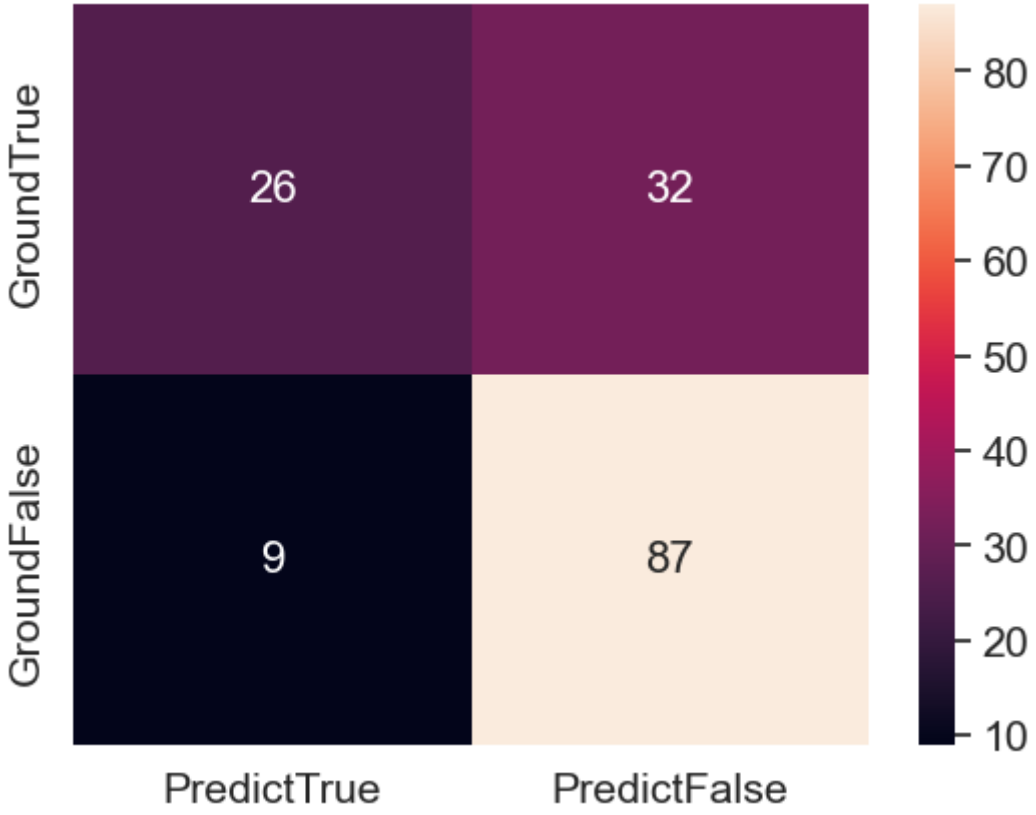
Parametry porzucone w momencie tworzenia macierzy pomyłek: BloodPressure, DPF.

Tym razem w przypadku błędu K2 brakuje jednego parametru na grafie, DPF a liczba powiązań wynosi 6. Bardzo silne powiązania występują pomiędzy Insulin -> Glucose oraz Glucose -> Outcome. Parametry, których wartości bezpośrednio zależą od obecności cukrzycy to BMI oraz Age. Obserwując tablice pomyłek, można zauważyć, że dokładność modelu wynosi 76,623%. Najmniejsze uzyskane prawdopodobieństwo wyniosło 50,025%, natomiast największe 87,165%.

W przypadku błędu BIC na grafie brakuje jednego parametru- DPF, natomiast BloodPressure nie jest powiązany z żadną inną wielkością. Liczba powiązań wynosi 5. Najsilniejsze powiązanie występuje pomiędzy Glucose -> Outcome. Parametry, których wartości są powiązane z obecnością cukrzycy to BMI, Insulin oraz Pregnancies. Dokładność modelu wynosi 75,325%. Najmniejsze uzyskane prawdopodobieństwo wyniosło 50,208%, natomiast największe 85,556%.

W przypadku błędu BDEU również na grafie brakuje jednego parametru- DPF, natomiast BloodPressure nie jest powiązany z żadną inną wielkością. Liczba powiązań wynosi 5. Silne powiązanie występuje pomiędzy Age -> Pregnancies. Dokładność modelu wynosi 76,623%. Minimalne prawdopodobieństwo wynosi 50,025%, a maksymalne 87,165%

Rysunek 22 Podział – histogramy, uzupełnienie danych – średnia arytmetyczna, błąd – K2

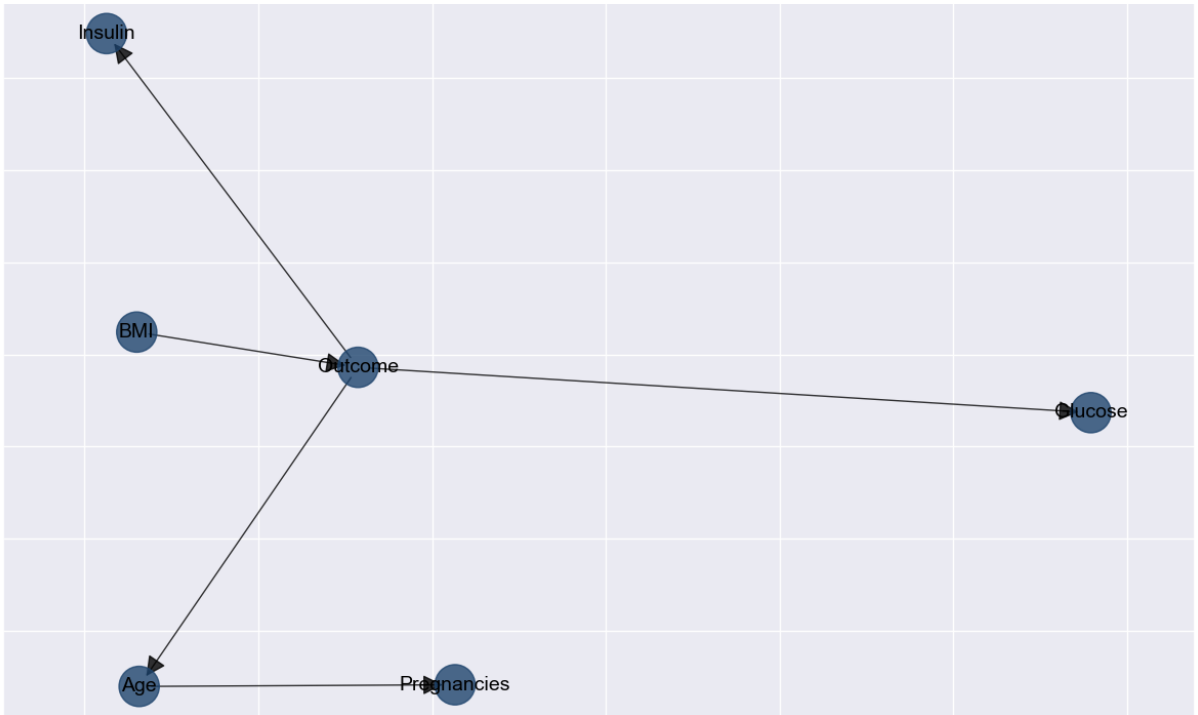


Tablica Pomyłek 10 Podział – histogramy, uzupełnienie danych – średnia arytmetyczna, błąd – K2

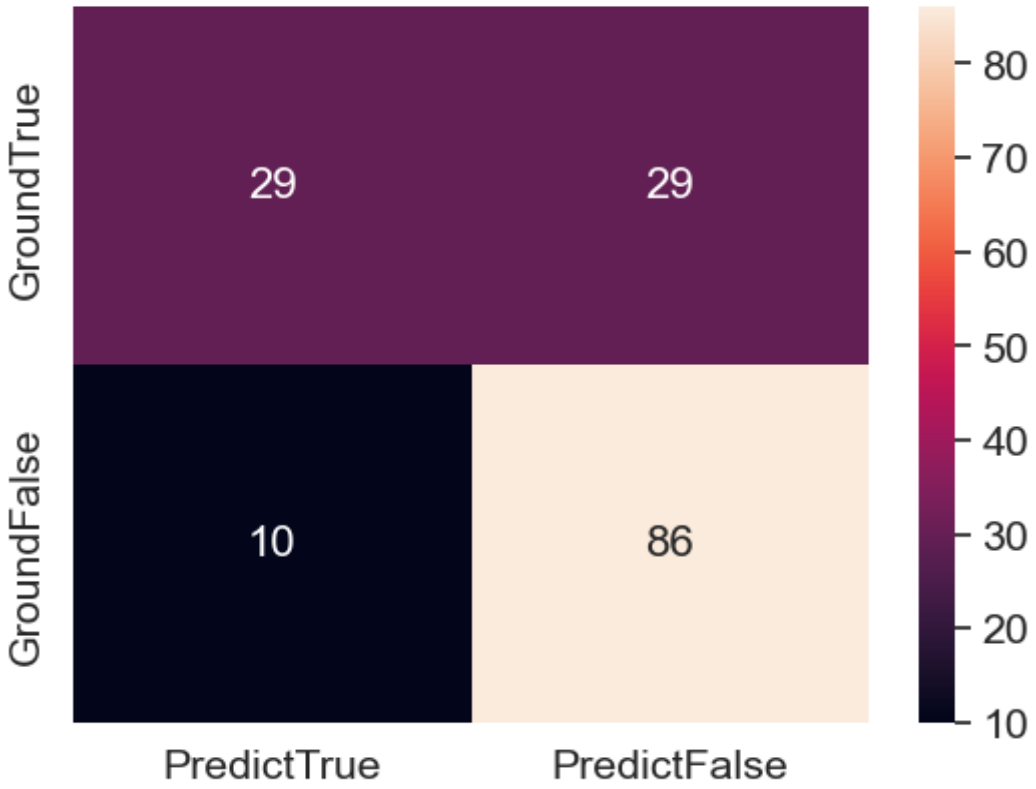
Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 26, Min. Prob. = 0.5210430414806857, Max. Prob. = 0.701127198120431
- FP = 9, Min. Prob. = 0.5098112300928379, Max. Prob. = 0.7011271981204311
- TN = 87, Min. Prob. = 0.5008134392348805, Max. Prob. = 0.8701199048256876
- FN = 32, Min. Prob. = 0.5008134392348805, Max. Prob. = 0.7914339884859638

Parametry porzucone w momencie tworzenia macierzy pomyłek: DPF.



Rysunek 23 Podział – histogramy, uzupełnienie danych – średnia arytmetyczna, błąd – BIC

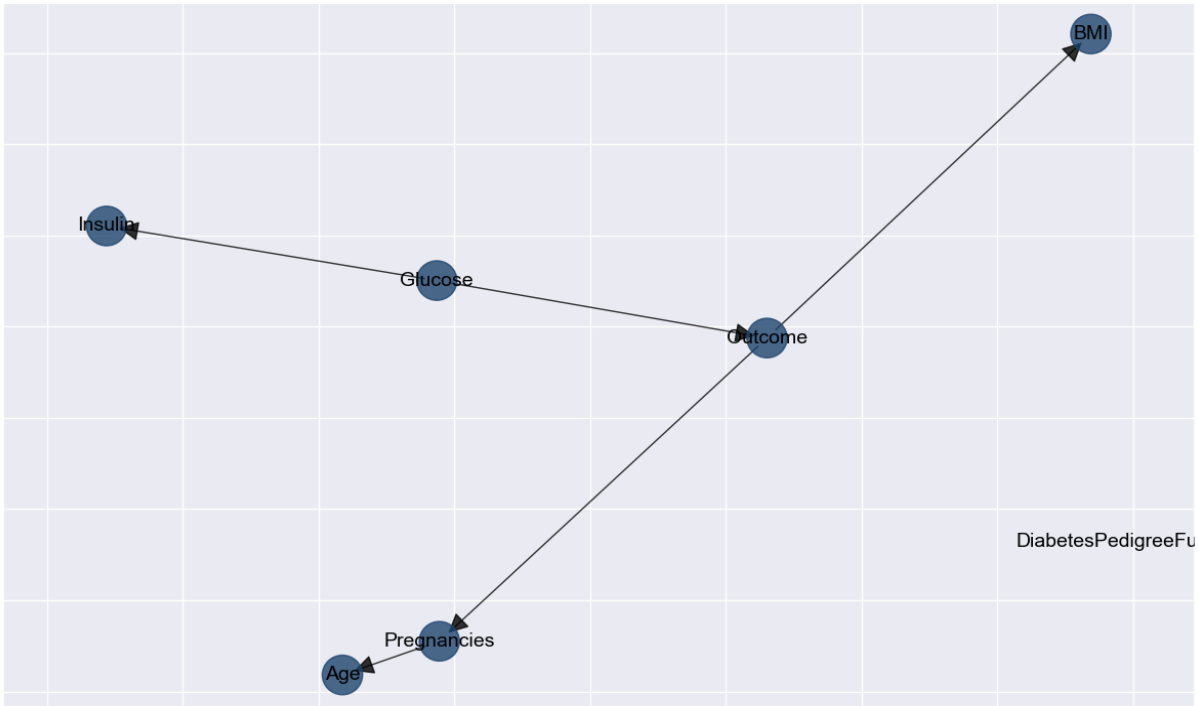


Tablica Pomyłek 11 Podział – histogramy, uzupełnienie danych – średnia arytmetyczna, błąd – BIC

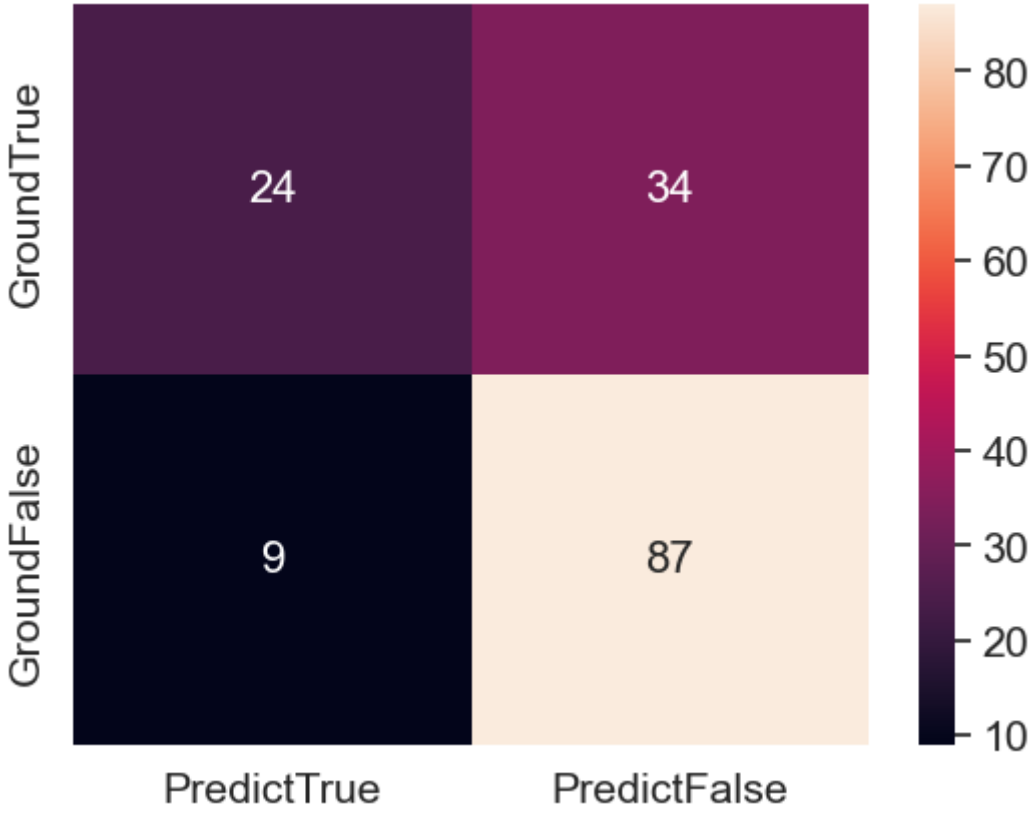
Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 29, Min. Prob. = 0.5068229790049557, Max. Prob. = 0.7263190221013869
- FP = 10, Min. Prob. = 0.5110158550438579, Max. Prob. = 0.6823949535517283
- TN = 86, Min. Prob. = 0.5023569057293481, Max. Prob. = 0.8957628631630161
- FN = 29, Min. Prob. = 0.5385662454925649, Max. Prob. = 0.8114117268422043

Parametry porzucone w momencie tworzenia macierzy pomyłek: DPF, BloodPressure.



Rysunek 24 Podział – histogramy, uzupełnienie danych – średnia arytmetyczna, błąd – BDEU



Tablica Pomyłek 12 Podział – histogramy, uzupełnienie danych – średnia arytmetyczna, błąd – BDEU

Minimalne i maksymalne wartości prawdopodobieństw:

- TP = 24, Min. Prob. = 0.5014186642922263, Max. Prob. = 0.6873721344837177
- FP = 9, Min. Prob. = 0.5133190709114234, Max. Prob. = 0.6420147883660722
- TN = 87, Min. Prob. = 0.5131720680954753, Max. Prob. = 0.8159393861470009
- FN = 34, Min. Prob. = 0.504330508734158, Max. Prob. = 0.7775609786090274

Parametry porzucone w momencie tworzenia macierzy pomyłek: DPF, BloodPressure.

W przypadku błędu K2 brakuje jednego parametru na grafie- DPF, a liczba powiązań wynosi 6. Bardzo silne powiązanie występuje pomiędzy BloodPressure -> Age. Parametry, których wartości bezpośrednio zależą od obecności cukrzycy to BMI oraz Glucose. Obserwując tablice pomyłek, można zauważyć, że dokładność modelu wynosi 73,377%. Najmniejsze uzyskane prawdopodobieństwo wyniosło 50,081%, natomiast największe 87,012%.

W przypadku błędu BIC na grafie brakuje dwóch parametrów- DPF oraz BloodPressure. Liczba powiązań wynosi 5. Najsilniejsze powiązanie występuje pomiędzy BMI -> Outcome. Parametry, których wartości są powiązane z obecnością cukrzycy to Insulin, Glucose oraz Age. Dokładność modelu wynosi 74,675%. Najmniejsze uzyskane prawdopodobieństwo wyniosło 50,236%, natomiast największe 89,576%.

W przypadku błędu BDEU na grafie brakuje jednego parametru- BloodPressure, natomiast DPF nie jest powiązany z żadną inną wielkością. Liczba powiązań wynosi 5. Silne powiązanie występuje pomiędzy Pregnancies -> Age. Dokładność modelu wynosi 72,078%. Minimalne prawdopodobieństwo wynosi 50,141%, a maksymalne 81,594%.

Podsumowanie, obserwacje i wnioski:

Obserwując wyniki można zauważyć, że parametrami, które najczęściej nie znajdywały żadnego powiązania na grafie są DiabetesPedigreeFunction oraz BloodPressure. DiabetesPedigreeFunction to jedyny parametr, do którego nie uzyskaliśmy przedziałów

zaproponowanych przez specjalistę, a BloodPressure miał najwięcej (nie licząc SkinThickness, które zostało całkowicie odrzucone) błędnych wyników, a co za tym idzie, najwięcej wartości „sztucznych" wyznaczonych poprzez regresje bayesowską lub średnią arytmetyczną. Model o największej dokładności (80,519%) okazał się pierwszy model, a więc przedziały zaproponowane przez lekarza, uzupełnienie danych regresją bayesowską oraz błąd K2. Wszystkie modele z grupy lekarskiej i zastosowanej regresji bayesowkiej okazały się lepsze średnio o około 5 punktów procentowych od innych modeli. Wykorzystanie wiedzy specjalistycznej zapewniło ustawienie przedziałów w miejscach najważniejszych z punktu widzenia diagnozy pacjenta, natomiast wykorzystanie regresji bayesowskiej to uzupełnienia wadliwych danych pomogło w zachowaniu większej różnorodności danych (w przypadku średniej arytmetycznej wszystkie uzupełnione wartości danego parametru miały tą samą liczbę). Uzyskana dokładność modeli jest bardzo zadowalająca, ponieważ diagnoza człowieka w kontekście medycznym jest bardzo trudnym i złożonym procesem, w którym często nawet specjaliści popełniają błędy. W przypadku zaprezentowanym w tym raporcie program z założenia ma przydzielać odpowiedni priorytet pacjentowi lub sugerować diagnozę dla danego pacjenta, a nie być podmiotem decyzyjnym, dlatego wynik rzędu 70%-80% jest satysfakcjonujący.

Sieci Bayesowskie okazały się skutecznym algorytmem wspomagania decyzji w problemie klasyfikacji pacjenta pod kątem choroby na cukrzyce. Pomimo niewielkiego zbioru (liczba pacjentów nie przekraczała 1000 osób) udało się uzyskać zadowalające wyniki końcowe. W przyszłości w celu rozszerzenia tego projektu można spróbować wykorzystać bardziej rozbudowany zbiór zawierające dane pacjentów z całego świata (nie tylko z Indii) oraz każdej płci (nie tylko kobiecej). Nadal dużym problemem okazują się sprawy związane ściśle z medycyną (podział zbioru na podzbiory), gdzie podejście analityczne ustępuje wiedzy specjalistycznej. Korzystając z opinii lekarza znacząco przyspieszono prace nad samym projektem, a wyniki modelu znacząco się poprawiły.