# Diabetes

Zápočtový projekt ISR • Bc. Filip Procházka

# Vytyčené cíle

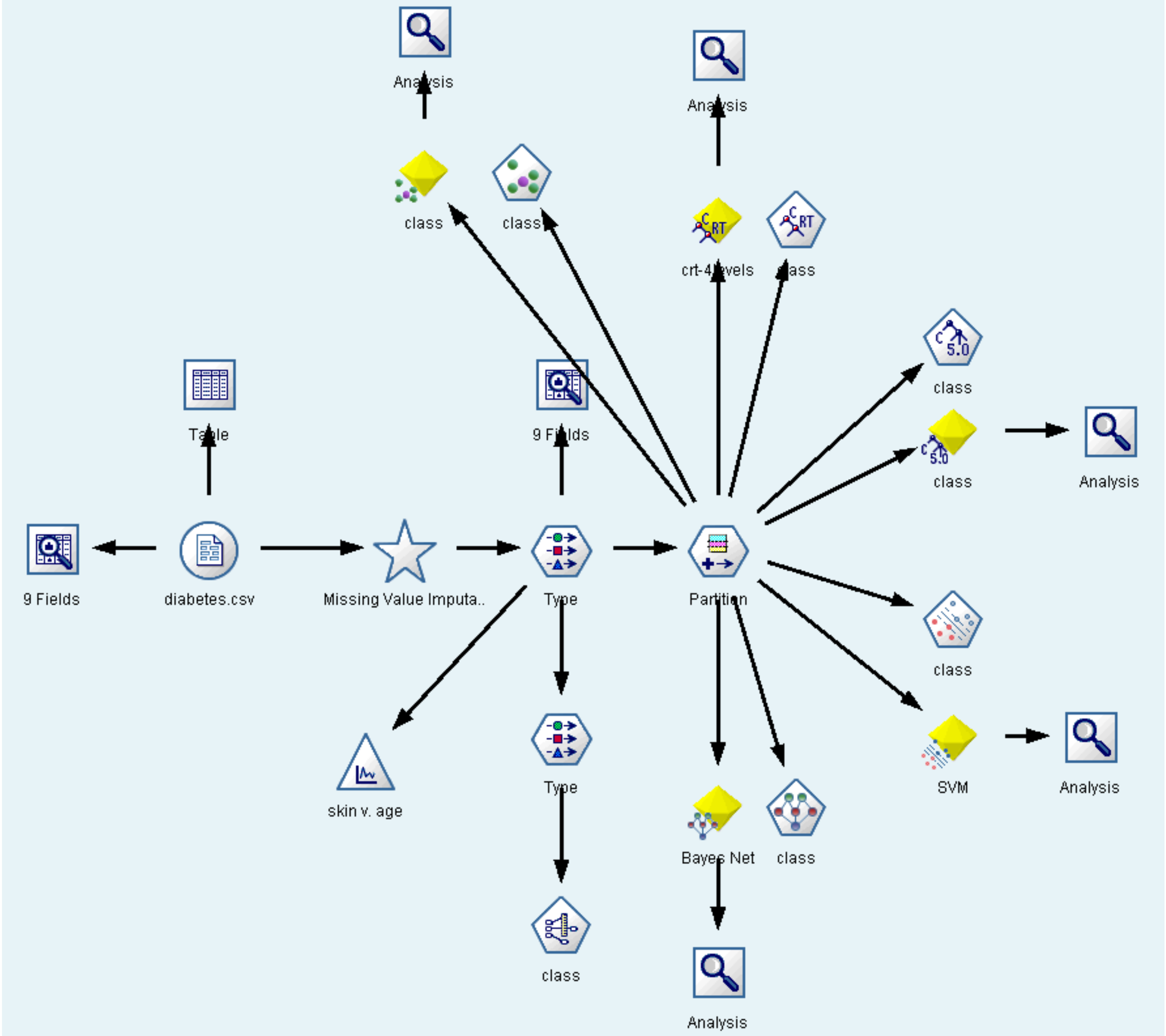
- **Diagnostika diabetu:**
  - 1. **Maximalizace** počtu pozitivně diagnostikovaných.
  - 2. **Minimalizace** chybné predikce u nemocných.

# Kvalita a úpravy dat

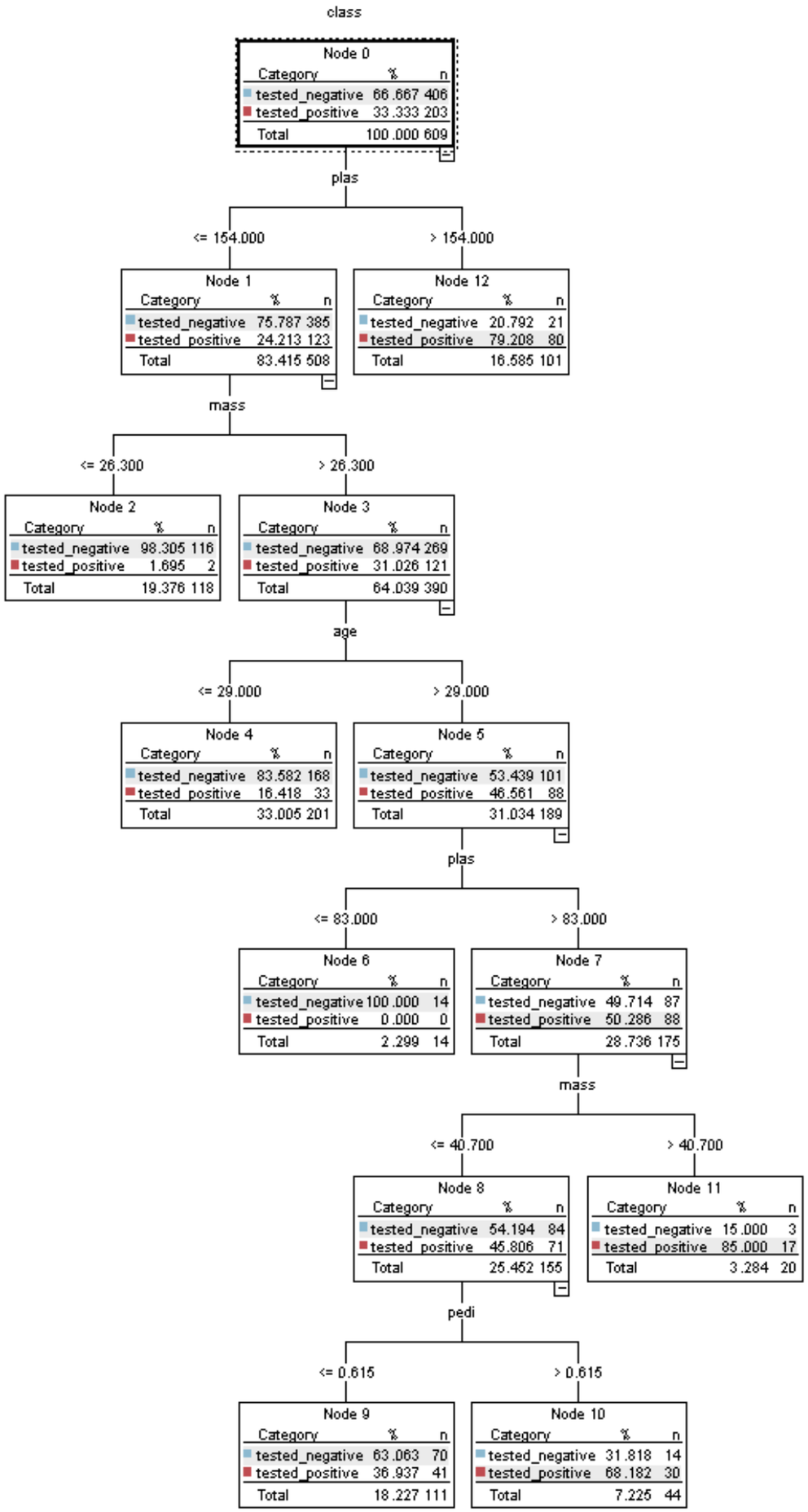
- 768 instancí.
- **7 vstupních** atributů **číselného** typu.
- **1 výstupní** atribut typu **flag.**
- Chybějící hodnoty reprezentované hodnotou 0.
- Nahrazení chybějících hodnot mediánem.
- Rozdělení na **trénovací** a **testovací** (**80:20**)

# Data

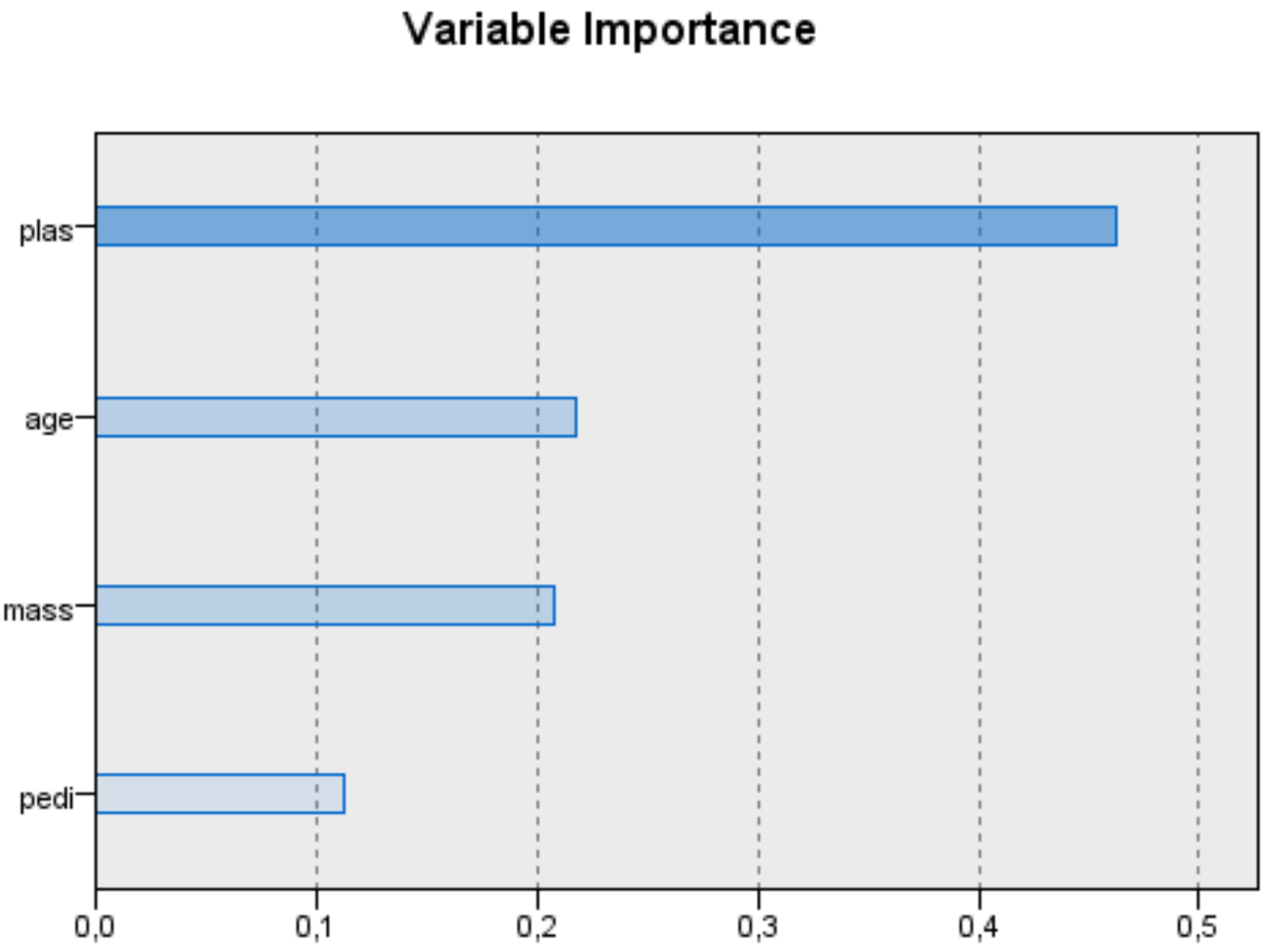
| název atributu | vstupní/výstupní | význam |
|---|---|---|
| preg | vstupní | počet těhotenství v životě |
| plas | vstupní | koncentrace plazmatické glukózy v plazmě žilní krve po 2 hodinách od orálního podání glukózy (mmol/l) |
| pres | vstupní | diastolický krevní tlak (mmHg) |
| skin | vstupní | tloušťka kožní řasy nad tricepsem (mm) |
| insu | vstupní | hladina inzulinu v séru (mU/l) |
| mass | vstupní | BMI |
| pedi | vstupní | diabetes mellitus pedigree function |
| age | vstupní | věk (let) |
| class | výstupní | onemocnění diabetem (tested_positive/ tested_negative) |


Analysis
Analysis
class
class
crt-4 levels
class
Table
9 Fields
class
Analysis
9 Fields
diabetes.csv
Missing Value Imputa..
Type
Partition
class
skin v. age
Type
SVM
Analysis
Bayes Net
class
class
Analysis

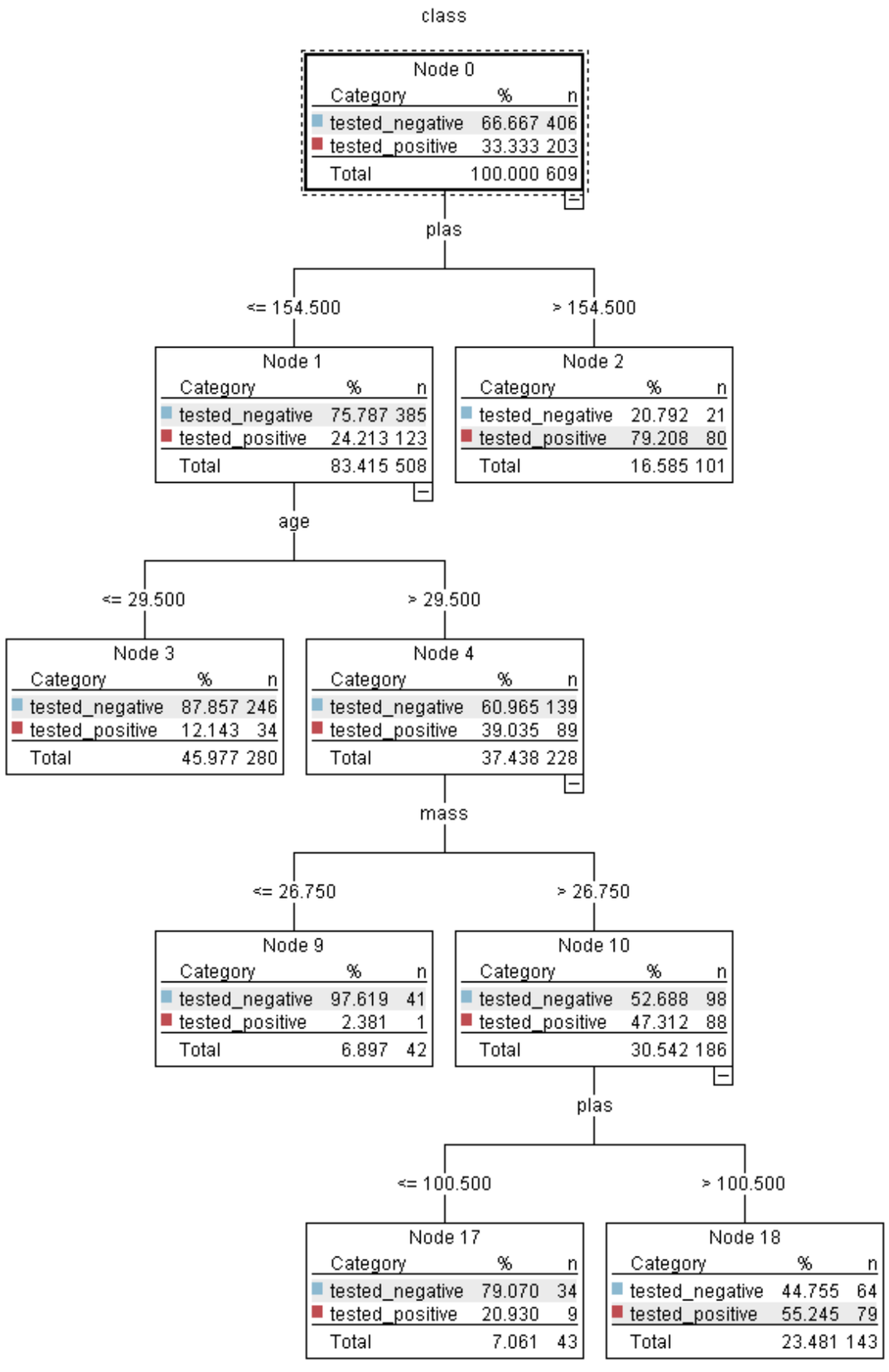
# C5.0


class
Node 0
Category % n
tested_negative 66.667 406
tested_positive 33.333 203
Total 100.000 609
plas
<= 154.000
> 154.000
Node 1
Category % n
tested_negative 75.787 385
tested_positive 24.213 123
Total 83.415 508
Node 12
Category % n
tested_negative 20.792 21
tested_positive 79.208 80
Total 16.585 101
mass
<= 26.300
> 26.300
Node 2
Category % n
tested_negative 98.305 116
tested_positive 1.695 2
Total 19.376 118
Node 3
Category % n
tested_negative 68.974 269
tested_positive 31.026 121
Total 64.039 390
age
<= 29.000
> 29.000
Node 4
Category % n
tested_negative 83.582 168
tested_positive 16.418 33
Total 33.005 201
Node 5
Category % n
tested_negative 53.439 101
tested_positive 46.561 88
Total 31.034 189
plas
<= 83.000
> 83.000
Node 6
Category % n
tested_negative 100.000 14
tested_positive 0.000 0
Total 2.299 14
Node 7
Category % n
tested_negative 49.714 87
tested_positive 50.286 88
Total 28.736 175
mass
<= 40.700
> 40.700
Node 8
Category % n
tested_negative 54.194 84
tested_positive 45.806 71
Total 25.452 155
Node 11
Category % n
tested_negative 15.000 3
tested_positive 85.000 17
Total 3.284 20
pedi
<= 0.615
> 0.615
Node 9
Category % n
tested_negative 63.063 70
tested_positive 36.937 41
Total 18.227 111
Node 10
Category % n
tested_negative 31.818 14
tested_positive 68.182 30
Total 7.225 44

# C5.0

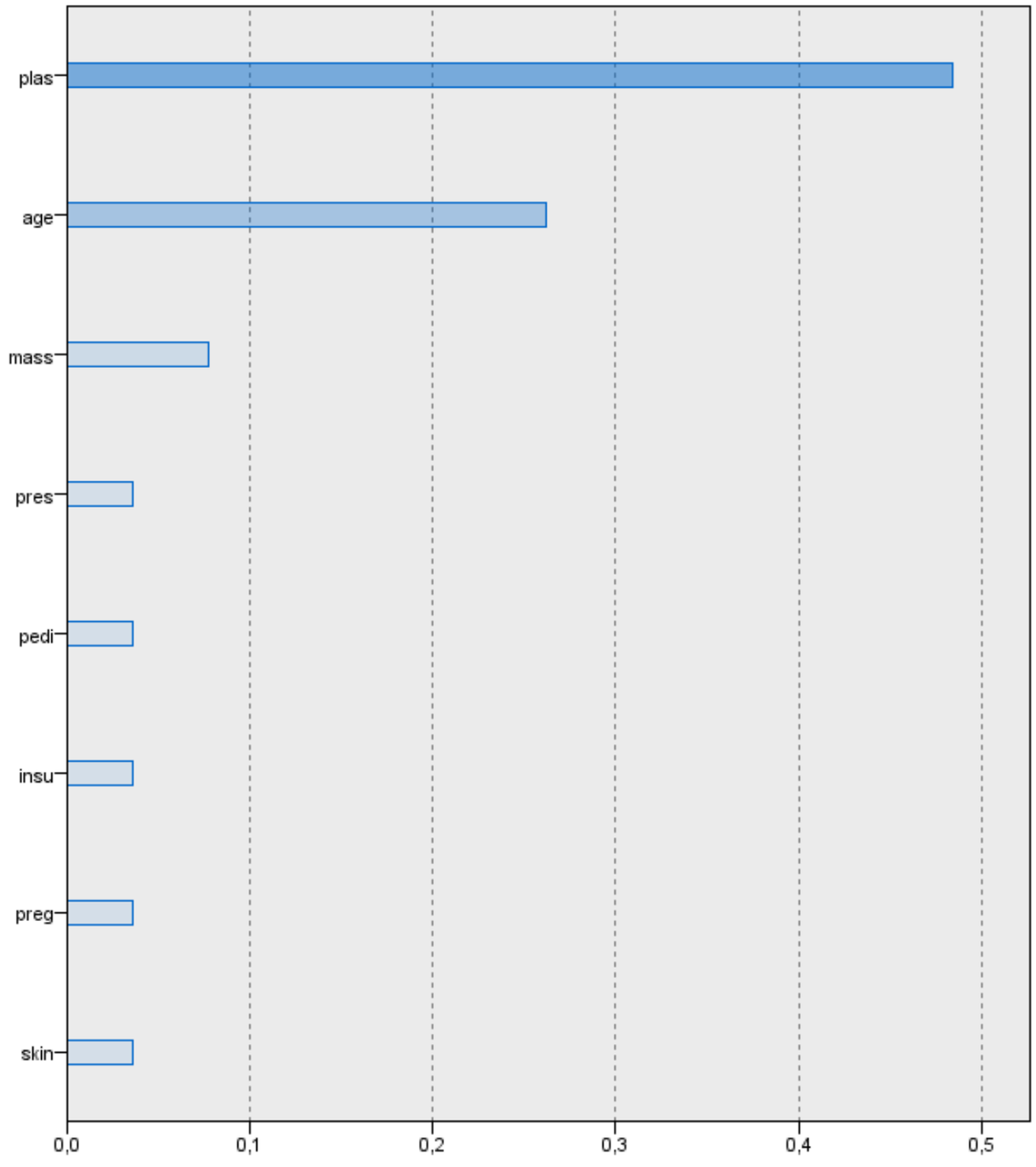


# CRT


class
Node 0
Category % n
tested_negative 66.667 406
tested_positive 33.333 203
Total 100.000 609
plas
<= 154.500
> 154.500
Node 1
Category % n
tested_negative 75.787 385
tested_positive 24.213 123
Total 83.415 508
Node 2
Category % n
tested_negative 20.792 21
tested_positive 79.208 80
Total 16.585 101
age
<= 29.500
> 29.500
Node 3
Category % n
tested_negative 87.857 246
tested_positive 12.143 34
Total 45.977 280
Node 4
Category % n
tested_negative 60.965 139
tested_positive 39.035 89
Total 37.438 228
mass
<= 26.750
> 26.750
Node 9
Category % n
tested_negative 97.619 41
tested_positive 2.381 1
Total 6.897 42
Node 10
Category % n
tested_negative 52.688 98
tested_positive 47.312 88
Total 30.542 186
plas
<= 100.500
> 100.500
Node 17
Category % n
tested_negative 79.070 34
tested_positive 20.930 9
Total 7.061 43
Node 18
Category % n
tested_negative 44.755 64
tested_positive 55.245 79
Total 23.481 143

# CRT

Variable Importance



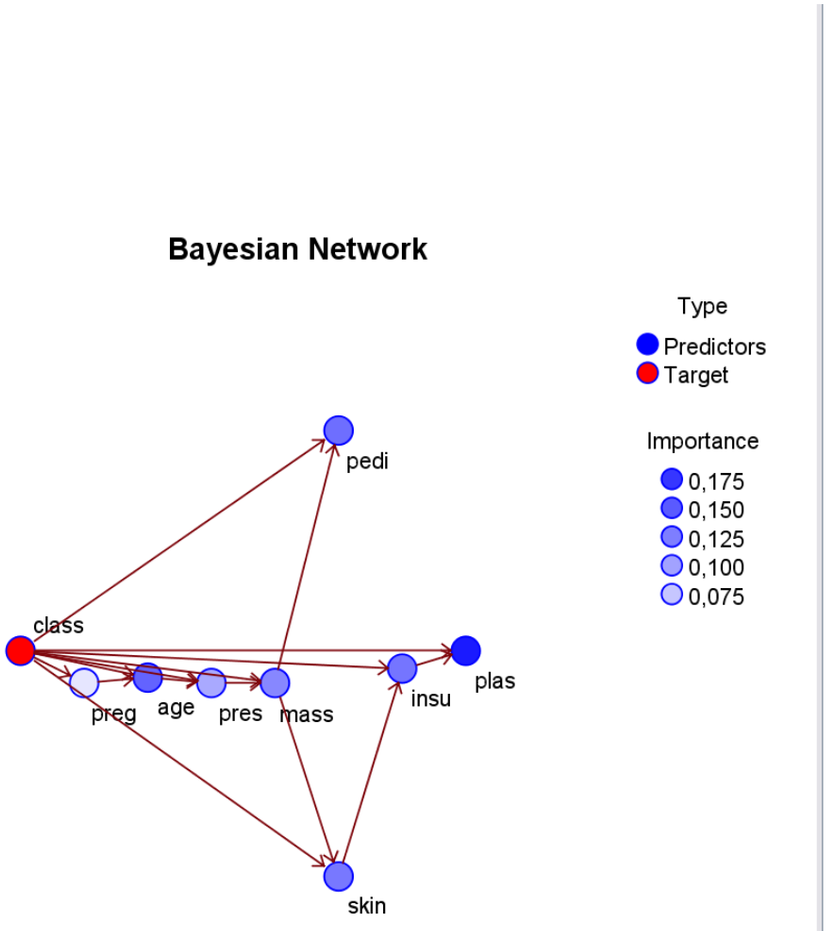
# GRI

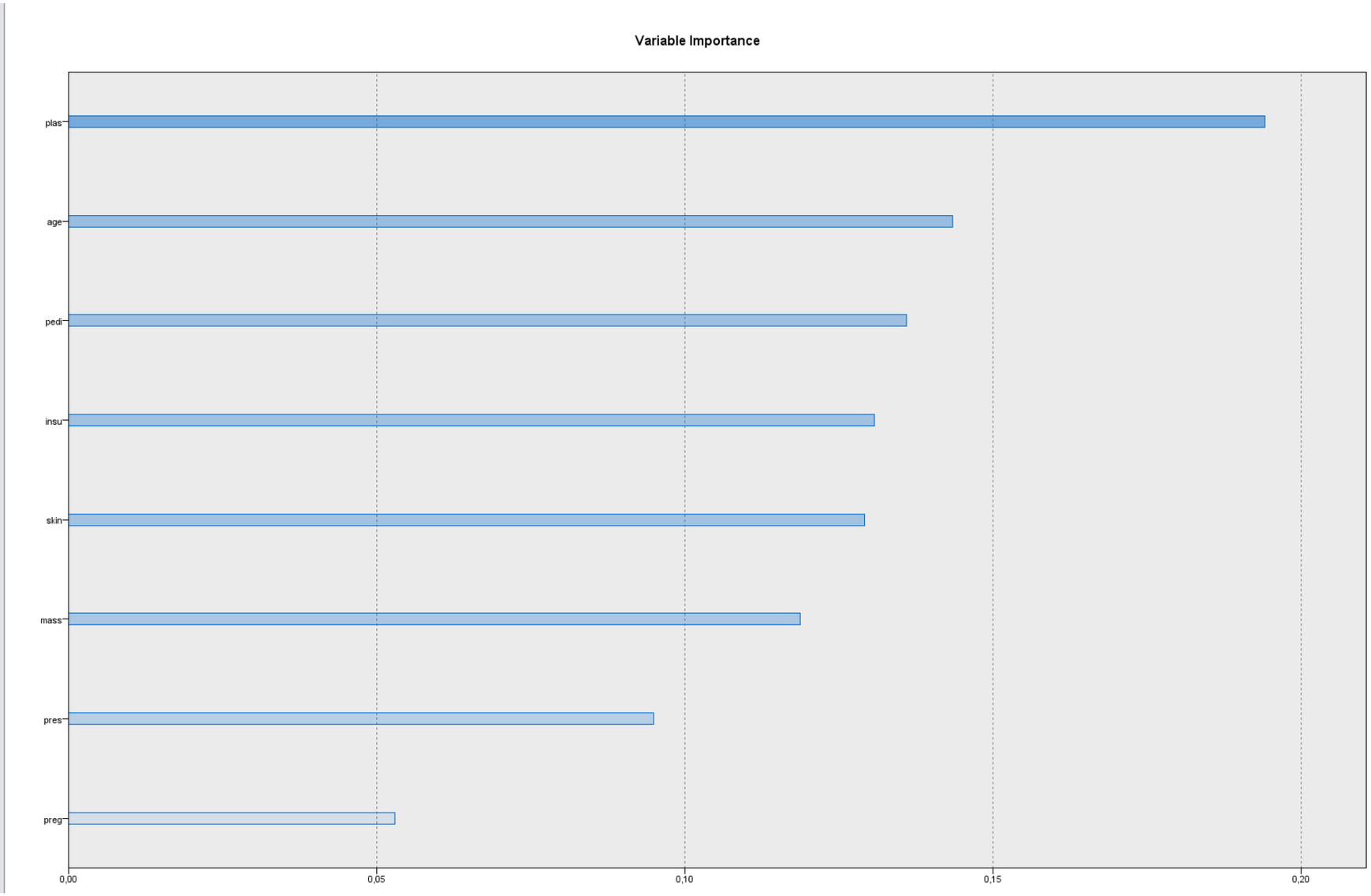
- Nejvýznamnější atributy:
  - **preg**, **plas**, **skin**, **mass**, **age**, **pedi**.
- Atributy **pres** a **insu** se vůbec nevyskytují.

Sort by: Support % | 16 of 16

| Consequent | Antecedent | Support % | Confidence % |
|---|---|---|---|
| class = tested_negative | preg > 4,500 | 35,94 | 52,17 |
| class = tested_negative | plas < 99,500 | 25,0 | 92,71 |
| class = tested_positive | preg > 6,500 | 22,01 | 56,21 |
| class = tested_negative | plas < 99,500<br>skin < 30,500 | 19,27 | 95,95 |
| class = tested_negative | mass < 26,350<br>plas < 129,500 | 16,02 | 99,19 |
| class = tested_positive | plas > 154,500 | 15,89 | 80,33 |
| class = tested_negative | age < 24,500<br>plas < 130,500<br>mass < 32,250 | 15,36 | 98,31 |
| class = tested_positive | preg > 6,500<br>mass > 27,850<br>plas > 106,500 | 12,24 | 78,72 |
| class = tested_positive | preg > 6,500<br>plas > 142,500 | 7,55 | 84,48 |
| class = tested_positive | preg > 6,500<br>preg < 9,500<br>plas > 139,500 | 6,12 | 87,23 |
| class = tested_positive | preg > 6,500<br>pedi > 0,254<br>plas > 142,500 | 5,86 | 91,11 |
| class = tested_positive | preg > 6,500<br>plas > 142,500<br>pedi > 0,254 | 5,86 | 91,11 |
| class = tested_negative | preg > 4,500<br>pedi < 0,672<br>plas < 96,500 | 4,69 | 97,22 |
| class = tested_negative | preg > 4,500<br>plas < 96,500<br>pedi < 0,596 | 4,43 | 97,06 |
| class = tested_negative | preg > 4,500<br>plas < 96,500<br>preg < 9,500 | 4,3 | 96,97 |
| class = tested_negative | preg > 4,500<br>preg < 9,500<br>plas < 96,500 | 4,3 | 96,97 |

# Bayes Net

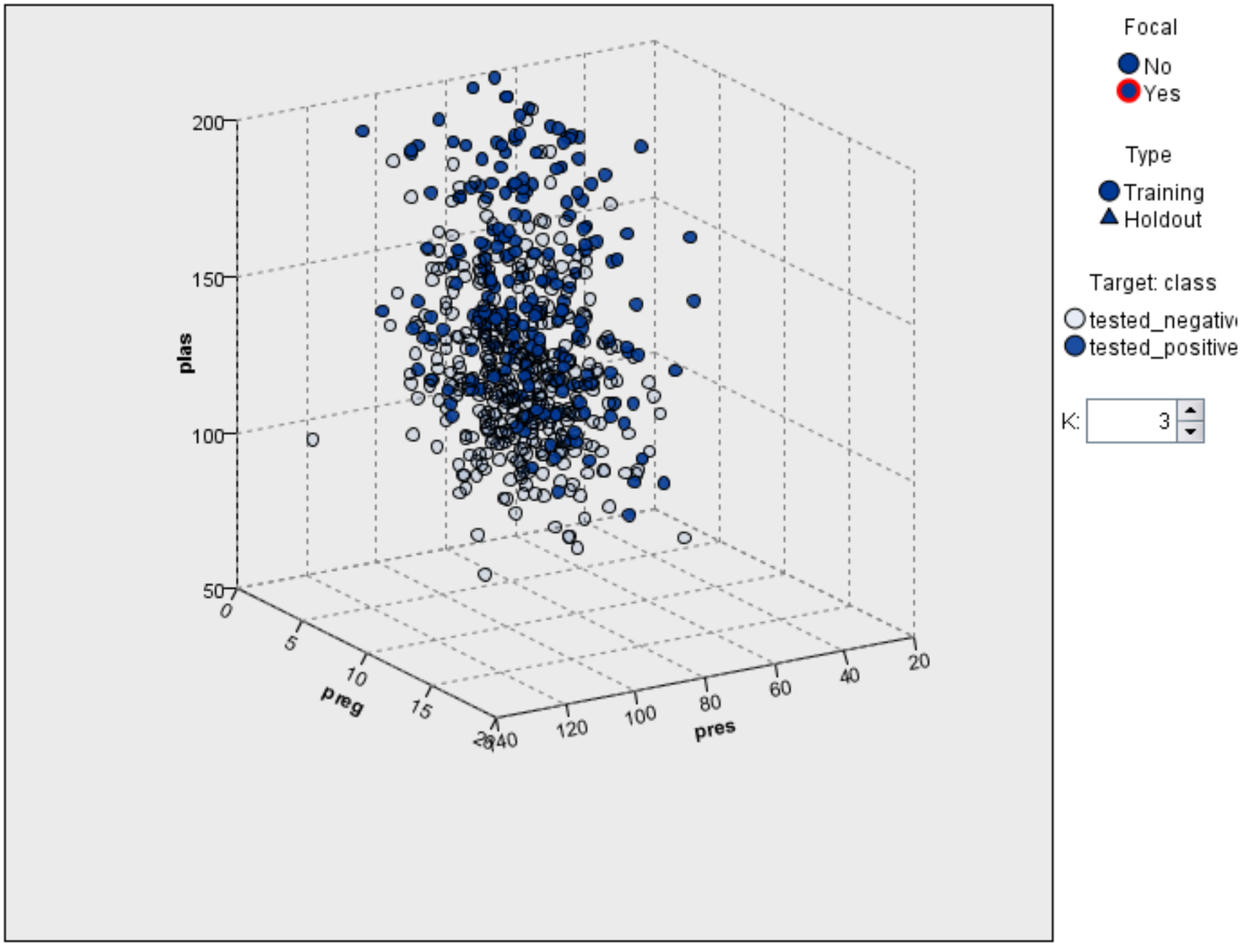

Bayesian Network
Type
Predictors
Target
Importance
0,175
0,150
0,125
0,100
0,075
class
pedi
plas
preg
age
pres
mass
insu
skin


Variable Importance
plas
age
pedi
insu
skin
mass
pres
preg
0,00
0,05
0,10
0,15
0,20

# KNN

**Feature Space**

**Built Model: 3 selected features, K = 3**




**Select points to use as focal cases**

**This chart is a lower-dimensional projection of the feature space,which contains a total of 8 features.**

# Nejvýznamnější atributy

- Nejvýznamnější atributy: **plas**, **age**, **mass** a **pedi**.
- Tzn.:
  - koncentrace plazmatické glukózy,
  - věk,
  - BMI,
  - diabetes melltius pedigree function.

# Závěr

Minimalizace chybné predikce u nemocných.

- Nejlepších výsledků dosáhl algoritmus **CRT.**
  - Na **trénovacích** i **testovacích** datech.
- **Úspěšnost** odhalení nemocných osob:
  - **78 %** na trénovacích datech,
  - **65 %** na testovacích datech.

# Závěr

Maximalizace počtu pozitivně diagnostikovaných

- Opět algoritmus **CRT.**
  - **Riziko**, že nemoc **nebude** odhalena:
    - **7 %** na trénovacích datech,
    - **14 %** na testovacích datech.
- U ostatních algoritmů přenost cca 20–24 %.

- Results for output field class
  - Comparing $R-class with class

| 'Partition' | 1_Training | | 2_Testing | |
|---|---|---|---|---|
| Correct | 480 | 78,82% | 118 | 74,21% |
| Wrong | 129 | 21,18% | 41 | 25,79% |
| Total | 609 | | 159 | |

  - Coincidence Matrix for $R-class (rows show actuals)

| 'Partition' = 1_Training | tested_negative | tested_positive |
|---|---|---|
| tested_negative | 321 | 85 |
| tested_positive | 44 | 159 |

| 'Partition' = 2_Testing | tested_negative | tested_positive |
|---|---|---|
| tested_negative | 76 | 18 |
| tested_positive | 23 | 42 |

  - Performance Evaluation

| 'Partition' = 1_Training | |
|---|---|
| tested_negative | 0,277 |
| tested_positive | 0,67 |

| 'Partition' = 2_Testing | |
|---|---|
| tested_negative | 0,261 |
| tested_positive | 0,538 |

  - Confidence Values Report for $RC-class

| 'Partition' = 1_Training | |
|---|---|
| Range | 0,552 - 0,955 |
| Mean Correct | 0,807 |
| Mean Incorrect | 0,694 |
| Always Correct Above | 0,955 (0% of cases) |
| Always Incorrect Below | 0,552 (0% of cases) |
| 97,62% Accuracy Above | 0,876 |
| 2,0 Fold Correct Above | 0,976 (87,58% of cases) |

| 'Partition' = 2_Testing | |
|---|---|
| Range | 0,552 - 0,955 |
| Mean Correct | 0,793 |
| Mean Incorrect | 0,746 |
| Always Correct Above | 0,876 (4,4% of cases) |
| Always Incorrect Below | 0,552 (0% of cases) |
| 100% Accuracy Above | 0,876 |
| 2,0 Fold Correct Above | 1,0 (87,58% of cases) |

# Závěr

- Pro účely primární lékařské diagnostiky přesnost nedostačuje.

- Pro nelékařské účely však může dostačovat.